Young learners' processing of multimodal input and its impact on reading comprehension: An eye-tracking study

Ana Pellicer-Sánchez[1], Elsa Tragant[2], Kathy Conklin[3], Michael Rodgers[4], Raquel Serrano[2], Angels Llanes[5]

[1]UCL Institute of Education, [2]University of Barcelona, [3]University of Nottingham, [4]Carleton University, [5]University of Lleida

ABSTRACT

Theories of multimedia learning suggest that learners can form better referential connections when verbal and visual materials are presented simultaneously. Furthermore, the addition of auditory input in reading-while-listening conditions benefits performance on a variety of linguistic tasks. However, little research has been conducted on the processing of multimedia input (written text and images) with and without accompanying audio. Eye movements were recorded during young L2 learners' (N = 30) processing of a multimedia story text in reading-only and reading-while-listening conditions in order to investigate looking patterns and their relationship with comprehension using a multiple-choice comprehension test. Analysis of the eye movement data showed that the presence of audio in reading-while-listening conditions allowed learners to look at the image more often. Processing time on text was related to lower levels of comprehension, whereas processing time on images was positively related to comprehension.

INTRODUCTION

Reading materials for young, English as foreign language (EFL) learners often come with pictures that illustrate and support the content of the text and that make the reading passages more engaging. Research has shown that pictures play a major role in the development of listening and reading skills, as they contribute to the creation of contexts that affect the meanings derived from words (Wright, 2010). The non-verbal information in pictures allows learners to predict what the text is about, making the construction of meaning easier, and helps them keep the overall context in mind as well as information about the characters in the text and the situations they are in (Wright, 2010).

The benefit of the simultaneous presentation of verbal and non-verbal information is supported by the multimedia learning hypothesis, which suggests that people learn more deeply from input that includes both words (written or spoken) and pictures (including both static graphs, pictures and dynamic videos or animations) than from words alone (Mayer, 2001, 2009, 2014). The verbal input in multimedia materials can be presented via different sensory modalities, i.e., visual and auditory. Materials that involve multiple sensory modalities (i.e., visual, auditory, and kinaesthetic) have also been referred to as multimodal learning environments (e.g., Massaro, 2012). Thus, multimodal learning involves learning from a combination of sensory modes, while multimedia learning refers to learning from text (written or spoken) and pictures. In the context of content learning and knowledge construction in a first language (L1), theories of multimedia learning and empirical investigations supporting those theories have suggested that presenting a text via auditory and written modes, when pictures are also present, results in redundant information that might be detrimental for learning and comprehension (e.g., Kalyuga & Sweller, 2014). However, ample evidence for the positive effect of this redundancy has been provided in the second language (L2) context. Many studies have shown the advantage of reading-while-listening

(RWL) conditions for the acquisition of a variety of linguistic components in an L2, including reading fluency, comprehension, and vocabulary learning (e.g., Brown, Waring, & Donkaewbua, 2008; Chang, 2009; Chang & Millet, 2014, 2015; Webb & Chang, 2015).

Despite the reported benefits of combining text and pictures in multimedia reading materials, very little is known about how learners process the different input sources in these learning conditions. In the first language (L1), studies have used eye-tracking to explore learners' online processing of text and pictures in the context of science and maths learning (e.g., Johnson & Mayer, 2012; Mason et al., 2013; Mason, Tornatora, & Pluchino, 2015) and have provided useful insights about how learners integrate the different input sources. Unfortunately, we do not have a clear picture yet of how verbal and non-verbal input sources are processed in the context of L2 learning in the presence of auditory input and, importantly, how potential processing differences might be related to learning and comprehension. In a recent exploratory study, Serrano and Pellicer-Sánchez (2019) showed that the presence of auditory input led to differences in the allocation of attention in an illustrated graded reader in an L2, with more looks to the pictures in the RWL condition than in a reading-only (RO) condition. Importantly, Serrano and Pellicer-Sánchez provided initial evidence for the relationship between processing patterns and comprehension, suggesting that longer processing times on the text in both RWL and RO reflected processing difficulties that were then related to lower comprehension scores. However, the authors call for more research on this topic, as their results could be due to the small set (N = 10) of rather challenging questions used in the study. In addition, despite the reported positive effect that pictures have on reading comprehension (e.g., Elley & Mangubhai, 1983; Omaggio, 1979), no previous studies have looked at the relationship between the processing of pictures and comprehension in both RO and RWL conditions. Thus, our understanding of the relationship between the allocation of attention to the different input sources in multimedia materials and

comprehension is rather limited. The present study addresses this gap by using eye-tracking to examine how young EFL learners (11-12-year olds with 5 years of English instruction) process text and pictures in RO and RWL conditions and the impact that the potential differences in the allocation of attention to both text and pictures has on comprehension.

BACKGROUND

Principles of multimedia and multimodal learning

The multimedia principle, put forward by Mayer (2001), states that people learn better from words and pictures than from words alone. Multimedia learning has been shown to lead not only to better learning outcomes, but also to higher levels of motivation for learners (e.g., Sung & Mayer, 2013). Forms of presentation in multimedia environments are categorised according to the presentation modes (i.e., pictorial and verbal presentation) and the sensory modalities (i.e., auditory and visual presentation) (Mayer, 2014). As Mayer (2014) explains, the presentation mode relates to Paivio's (1986, 2006) Dual Coding Theory, which suggests that the two modes (i.e., verbal and non-verbal) are processed through two different channels, each with limited processing capacity. The simultaneous activation of the verbal and non-verbal systems fosters learning. Notably, learning from a combination of sensory modalities (i.e., visual, auditory, kinaesthetic) is also referred to as multimodal learning (e.g., Massaro, 2012; Niegeman & Heidig, 2012).

Based on the available empirical evidence, Mayer (2009) identified twelve principles for the creation of effective multimedia learning environments. Two of those principles are particularly relevant for the present study, i.e., the *Redundancy principle* and the *Modality principle*. One of the most important principles of multimedia learning is the *Redundancy principle*, which suggests that redundant material (i.e., material that is concurrently presented in different forms or unnecessarily elaborated) interferes with learning (Kalyuga & Sweller,

2014). According to this principle, "people learn better from graphics and narration than from graphics, narration, and printed text" (Niegeman & Heidig, 2012, p. 2374). Duplication of the same information may overload working memory, inhibiting comprehension and learning (Kalyuga & Sweller, 2014). According to cognitive load theory (Sweller, 1988), the need to coordinate this redundant information involves a higher cognitive demand, which can have a detrimental effect on learning and comprehension (Kalyuga & Sweller, 2014). Interestingly, Kalyuga and Sweller (2014) claim that the negative effects of the simultaneous presentation of written and spoken text might be particularly evident in second or foreign language learning. However, empirical studies supporting the redundancy principle have mainly been conducted in the context of information acquisition and knowledge construction in an L1 (e.g., Kalyuga, Chandler, & Sweller, 1999; Jamet & Le Bohec, 2007; Mayer, Heiser, & Lonn, 2001). Similar evidence in the L2 context is scarce (e.g., Moussa-Inaty, Ayres, & Sweller, 2012) and contrasts with the positive effect of combining written and spoken texts found in RWL studies in the L2 (see review in the next section), as well as with the evidence provided by studies supporting the role of subtitles and captions on comprehension and L2 vocabulary learning (e.g., Montero Perez et al., 2014; Peters, 2019).

Also relevant for the present study is the *Modality principle*, which suggests that people learn better when pictures are presented with auditory text than with written text (Mayer & Moreno, 1998; Moreno & Mayer, 1999; Schnotz, 2014). According to the modality principle, the simultaneous presentation of written text and illustrations involves split attention which could negatively impact learning. Interestingly, Schnotz (2014) predicts a *reversed modality effect*, by which in certain situations, written text with illustrations might be better than spoken text. As Schnotz (2014) explains, written text allows learners to pause and re-read difficult passages and gives readers the opportunity to adapt their perceptual processing to

their needs. These opportunities to pause and re-read could be particularly useful for L2 learners.

It is important to note that the principles of multimedia learning were introduced to explain processing of multimodal and multimedia materials in L1 learning, specifically the learning of science and maths, where a complex integration of sources is needed and where the text and illustrations are specifically designed to teach content (e.g., a figure showing an engine and the teacher's oral explanation about how it functions). On the other hand, the type of multimedia materials that are often used in the L2 context serve a different purpose and require a different level of integration. For example, in a graded reader, like the one used in the present study, the content of the text can be understood without processing the pictures. Thus, it requires less complex integration of different information sources.

Reading-while-listening in a second language

Research on the effectiveness of combining auditory and visual modes in RWL in an L2 abounds. There is some empirical evidence questioning the effectiveness of RWL, suggesting that it has a detrimental effect on learning and comprehension. In Diao and Sweller's (2007) study, for example, EFL adult learners (first year university students) were asked to read two texts (2 experimental sessions) in one of two instructional conditions, i.e., RO or RWL. Participants were asked to read the text twice and to complete a comprehension recall test after the reading. Results of the study showed that RWL led to lower reading comprehension scores than RO.

Despite the negative evidence provided by Diao and Sweller (2007), the majority of investigations in the L2 context have suggested a positive effect of RWL. Studies conducted with adult learners have shown that RWL interventions led to improvements on a range of linguistics components, including listening fluency (e.g., Chang, 2009), vocabulary learning

(e.g., Webb & Chang, 2015; Webb & Chang, 2017; Webb et al., 2013), listening comprehension (e.g., Chang, 2009), and reading rates and reading comprehension (e.g., Chang & Millet, 2015), with RWL often showing an advantage over other modalities such as RO or listening only. Although the evidence is scant, a few studies have also shown the beneficial effects of RWL for young learners. Lightbown (1992) compared the effects of an extensive RWL instructional intervention to teacher-led instruction with primary school children and showed that RWL was at least as effective as teacher-led treatment for the acquisition of receptive and productive skills. In a follow-up study, Lightbown et al. (2002) found that after six years of the extensive RWL intervention, learners performed as well as comparison groups in receptive measures and in measures of oral production, although the approach was not as effective for written production. Similarly, Trofimovich, Lightbown, Harter, and Song (2009) showed positive effects of RWL for young learners' pronunciation accuracy.

In general, the studies suggest that RWL is not only beneficial for a range of L2 tasks, but also that learners have positive attitudes towards this mode of instruction, for both adult (e.g., Brown et al., 2008; Chang, 2009; Chang & Millet, 2014) and young learners (e.g., Lightbown, Halter, White, & Horst, 2002; Tragant, Muñoz, & Spada, 2016; Tragant & Vallbona, 2018).

Eye-tracking studies on multimodal input

Eye-tracking allows researchers to examine the cognitive effort involved in processing different types of stimuli (i.e., written/spoken verbal stimuli, as well as non-verbal, visual stimuli) (Pellicer-Sánchez & Conklin, 2020). It provides measures of different elements of the eye-movement record: *saccades*, i.e., the rapid movements of the eyes; *fixations*, i.e. when the eyes stop; as well as *regressions*, i.e., movements back in a text while reading. Eye-

tracking research has shown interesting differences in processing patterns for text and images/scenes (for a review of research see Conklin, Pellicer-Sánchez, & Carrol, 2018). Research has shown that average fixation duration on images (260-330ms) tends to be longer than fixation durations on text in silent reading (225-250 ms), because during scene perception useful information is gained from a fairly wide field of view (Rayner, 2009). Eye-tracking studies of reading have also shown that, when compared to adult readers, children have slower reading rates, more fixations, longer fixation durations, less skipping, and more saccades (Rayner, 1998; 2009; Whitford & Joanisse, 2018). Different processing patterns have been found for monolingual and bilingual children, with bilingual children having longer fixation durations and longer reading times than monolingual children when reading in their L1 (e.g., Whitford & Joanisse, 2018). Longer fixation durations, more saccades and a higher number of fixations have also been found when bilingual children read in their L2 than in their L1 (e.g., Whitford & Joanisse, 2018).

There has recently been a growing interest in the use of eye-tracking in the context of multimedia and multimodal learning, but there is still fairly little research (Alemdag & Cagiltay, 2018). The use of eye-tracking in multimedia learning overcomes many of the limitations imposed by self-report measures and allows for a direct indication of cognitive processing during multimedia learning (Mayer, 2017). Eye-tracking can demonstrate how learners integrate the different sources of input that are presented simultaneously and use this to explore their potential impact on performance measures (see Alemdag & Cagiltay, 2018, for a review of eye-tracking research in the domain of multimedia learning).

The majority of studies investigating eye movements during multimedia learning focused on science and maths learning in the L1 with the aim of providing empirical evidence for the principles of multimedia learning. Studies conducted with adult learners have demonstrated that presenting spoken versus written text alongside visuals, results in more

processing time on the visualizations in the former case, and more time spent reading than looking at the visualisations in the latter (e.g., Schmidt-Weigand, Kohnert, & Glowalla, 2010). Research has also shown that learning improves when text and pictures are presented close to each other, i.e., spatial contiguity principle (e.g., Johnson & Mayer, 2012). Studies conducted with young learners in the L1 have also demonstrated that the integration of text and pictures supports retention and the application of newly learned knowledge (Mason, Tornatora, & Pluchino, 2015). Young learners' attention to relevant pictures seems to be positively related to learning scores (e.g., Eitel, 2016), and a better integration of text and pictures is associated with enhanced performance (Mason, Tornatora, & Pluchino, 2015).

In the L2 context, eye-tracking studies on multimedia and multimodal materials have mainly been concerned with the processing of subtitled videos. Previous research conducted with adult learners has shown that both the animation and subtitles are processed and that learners process subtitles regardless of the language of the soundtrack and the language of the subtitles (e.g., Bisson, van Heuven, Conklin & Tunney, 2014). In general, irregular reading patterns (i.e., higher skipping rate, fewer fixations, longer latencies) have been shown in the processing of subtitles (e.g., d'Ydewalle & de Bruycker, 2007), with processing patterns differing by L1 background (e.g., Winke, Gass & Sydorenko, 2013) and by proficiency level (e.g., Muñoz, 2017). Some evidence for the relationship between subtitle reading and learning measures has also been provided in the L2 context (e.g., Montero Perez, Peters, & Desmet, 2015). Apart from the examination of subtitled videos, very few studies have used eye-tracking to examine adult learners' processing of text and static pictures in multimodal materials in the L2 context. In a recent study, Warren, Boers, Grimshaw, and Siyanova-Chanturia (2018) examined L2 adult learners' eye movements to different gloss types (i.e., text only, picture only, and text+picture). They found that the presence of pictures in multimodal glosses led to less attention paid to the text, although these processing differences

were not reflected in the general comprehension of the text. Bisson et al. (2015) examined L2 learners' eye movements when they learnt new words that were presented in an L2 auditory mode with written L1 translations and pictures. They found that the presence of pictures reduced attention to the translations and that the time spent processing pictures was positively related to learning gains.

Very few eye-tracking studies on multimedia learning have been conducted with young L2 learners. Similar to research with adult learners, most of the available studies have focused on the processing of subtitled videos. This research has demonstrated that younger learners also show irregular reading patterns (i.e., higher skipping rate, fewer fixations, longer latencies) when processing subtitles (e.g., d'Ydewalle & de Bruycker, 2007), and when compared to adult learners, they skip fewer subtitles and spend more time on them (e.g., Muñoz, 2017). When the use of dynamic images has been compared to static pictures, eye-tracking has shown more processing of the visuals in the dynamic condition (e.g., Tragant & Pellicer-Sánchez, 2019).

While these studies focused on the processing of subtitles, a recent exploratory study by Serrano and Pellicer-Sánchez (2019) examined young L2 learners' eye movements to text and pictures in an illustrated graded reader and found that the presence of auditory text led to more time spent processing the images. They also explored the relationship between processing of the text and reading comprehension and reported a negative correlation between processing time on the text and reading comprehension. However, despite the differences reported in the processing of pictures in the presence of auditory input, the study did not examine whether the processing of pictures was also related to comprehension. As argued earlier, given the central role that pictures play in reading comprehension, it would be important to examine, not only how they are processed but also how their processing is related to comprehension. In addition, as acknowledged by the authors, the negative

relationship between processing of text and comprehension reported in the study could be attributed to the relatively small and challenging set of comprehension questions. The present study uses eye-tracking to examine young learners' processing of text and pictures in an illustrated graded reader as well as the relationship between processing patterns (on both text and pictures) and comprehension.

THE STUDY

The aim of the present study was to examine young L2 learners' processing of visual text and images in the presence and absence of the auditory text, as well as the relationship that viewing patterns on both text and pictures had on comprehension. The following research questions were addressed:

1. Does the presence of auditory input affect young learners' allocation of attention to the text and pictures in multimodal reading conditions?
2. Is the amount of attention allocated to the text and images related to comprehension?

In order to address these questions, participants were asked to read a multimodal text in RO and RWL conditions while their eye movements were recorded and to complete a comprehension test. Eye movements to text and image areas in the two multimodal conditions were analysed, and this was related to their performance on the comprehension test. Based on the results of the study by Serrano and Pellicer-Sánchez (2019), it was hypothesised that the presence of audio would lead to differences in the amount of time allocated to text and pictures. Regarding the second research question, the available findings suggest a negative relationship between time spent processing the text and comprehension (Serrano & Pellicer-Sánchez, 2019). However, eye-tracking studies in other contexts have shown a positive relationship between processing time and performance measures (e.g.,

Godfroid, et al., 2018; Pellicer-Sánchez, 2016). Thus, it was hypothesised that the relationship between processing time and comprehension could go in either direction. Although no previous studies have looked at the relationship between the processing of pictures and reading comprehension, there is evidence showing a positive relationship between processing of pictures and learning gains (e.g., Bisson et al., 2015). Based on these findings, a positive relationship between processing of pictures and comprehension was hypothesised.

METHODOLOGY

Participants

Participants in this study were 30 EFL Catalan-Spanish bilinguals in a primary school in Barcelona (Spain). They were all in grade 6 and their ages ranged from 11 to 12. All participants had the expected level of L1 literacy for their age group (as reported by the class teacher). They all had received 5 years of English instruction and their proficiency level was A1.1 according to the Common European Framework of Reference for Languages (CEFR). Prior to the experiment, participants' vocabulary knowledge was assessed using the X-Lex vocabulary size test (Meara & Milton, 2003). Results revealed that all participants had a mean vocabulary size between 1K and 2K (max = 2,600 words, min. = 1,100, M = 1,985, SD = 443). Data from two participants were removed from the analysis because they failed to complete one of the measurement instruments. Data from 28 learners (14 male, 14 female) were included in the analyses.

Materials

Reading materials

The graded reader 'The Canterville Ghost' (Wilde, 2012) (level A1.1, 300 headwords) was modified for the purposes of our study. The text from this graded reader was shortened to fit

the length of the experiment and some of the lower frequency words were deleted to ensure that the vocabulary included in the story was within learners' level of proficiency. For example, the words *pumpkin* and *crayon*, which have a lower frequency (8K and 9K), were deleted from the original story. The final version of the text had 566 words, 94.2% of which were within the first 1000 most frequent words (lexical profile of the text analysed with Lextutor, Cobb, n.d). Our aim was to have 95-98% of the words in the text from the 1K, as participants in the study (with a mean vocabulary size of 1,985) were likely to know the words in this level,[1] and this would indicate adequate comprehension of the text (Hu & Nation, 2000). One of the most frequent words in the text, i.e., *ghost,* was from the 2K level but, given its centrality in the narrative, we confirmed knowledge of this word with each participant before starting the reading activity. Thus, counting *ghost* as a known word meant a lexical coverage of 96.5%.

The text was presented across 14 pages, which constituted the 14 screens/trials in the eye-tracking experiment (see Appendix for a sample of the experimental trials). Fourteen images were selected from the original graded reader to be presented alongside the text. The selected images accompanied the same part of the text as in the original graded reader. The text and image stimuli were designed to control for many of the factors that are known to affect eye-movement behaviour. We wanted to ensure that the text that appeared on each page of the reading experiment had the same or very similar number of words and appeared in the same format (same font size and font style). The size of the original images was modified so that all of the images had the same size. The position of text and images was counterbalanced so that both types of stimuli appeared at the right and left of the display. The design of the illustrated story followed the spatial and contiguity principle of the Cognitive Theory of Multimedia Learning, which suggests that people learn better when words and

pictures are presented near to each other and simultaneously, rather than successively (Mayer, 2009).

Finally, the auditory stimuli for the RWL condition was recorded by a native speaker of British English for each page of text at a speech rate of 113 words per minute (wpm), similar to the rate of speech of the original audio provided by the publishers for this and other graded readers.[2]

Comprehension test

Since we wanted to explore the relationship between processing of text and pictures and comprehension, two types of questions were created: questions that could be answered by reading the text and questions that could only be answered by extracting information from the pictures. As explained earlier, pictures help readers to predict the content of the text, to keep information about the overall context and characters in mind, and facilitate meaning construction (Wright, 2010). In this sense, the images could also be helpful in answering the text-related questions, as they supported the content of the text. However, the image-related questions focused on specific visual features that were not reported in the text. Thus, we will refer to two types of questions: text+image questions and image-only questions. The narrative was first parsed into idea units (i.e., distinct events or actions that occurred in the course of the story) and these were then used to create multiple-choice questions. Each test item provided three options and a fourth 'I don't know'. The test was in Catalan to ensure comprehension of the content. Questions that related to the images could only be answered by having looked at the pictures (e.g., what a character was wearing, where a scene took place, etc.). A battery of 26 multiple-choice items was piloted prior to the experiment with a group of learners of similar characteristics (N = 46). The results of the pilot allowed us to examine the quality of the items in terms of discrimination and level of difficulty. Based on

these analyses 18 of the 26 questions were kept unchanged and 10 new questions were created. The final test included a total of 28 items, i.e., 19 text+image questions and 9 image-only questions. After administration of the final test, the level of difficulty and discrimination was checked again. Only three items in the final test had a low level of discrimination and they were discarded from analysis. Consequently, all analyses in the study are based on written responses to 25 multiple-choice items (16 text+image questions and 9 image-only items) (Cronbach's alpha = .80).

Procedure and Analysis

Data were collected individually in a quiet room in the participants' school. Instructions were provided orally in the children's L1. Participants were then asked to read the story for comprehension while their eye movements were recorded. The experiment followed a within-subjects design, with all participants being exposed to both RO and RWL conditions. Half of the story was presented in RO and the other half in RWL in a counterbalanced design. Participants were told about the existence of the two conditions and that they would have to answer some comprehension questions after reading the story. Although the audio was only included in one part of the story, participants were asked to wear the headphones for the duration of the story as it aided concentration and also helped to isolate any potential noise.

The story was presented on a 1280x1024 monitor and displayed over 14 screens. In the RO condition pages advanced with a mouse click, whereas in the RWL condition the pages advanced automatically when the audio recording finished. Eye movements were recorded with Tobii T120 at a sampling rate of 120Hz that has a typical accuracy of 0.5° (measured in ideal conditions) and 0.2° resolution. A five-point calibration and validation procedure was performed at the beginning of the experiment. No other calibrations were performed during the experiment. After the reading activity participants were asked to complete the

comprehension test with no time pressure. The reading task lasted around 20 minutes and the whole procedure around 50 minutes.

For the analysis of eye movements, two regions of interest were defined for each trial, surrounding the image and the block of text. Fixations shorter than 80 ms were removed from the dataset (1% of the data). The following eye-movement measures were extracted and analysed as measures of attention allocation:

- Dwell time % (the percentage of the sum of all fixation durations within each region of interest)

- Fixation count % (the percentage of the total number of fixations in a trial within each region of interest)

- Average fixation duration within each region of interest

A dichotomous scoring system was used to score the comprehension test (1 for correct responses and 0 for incorrect responses). In response to the first research question, we examined the effect of two independent variables, i.e., condition (RWL and RO) and region (text and picture) on the dependent variables, i.e., the three eye-movement measures, via linear mixed-effect models using the lme4 (v 1.1-21; Bates, Maechler, Bolker & Walker, 2015) package for R (v 3.6.1; R Core Team, 2019). The *p* values for the effects were obtained using the lmerTest package (3.1-0; Kuznetsova, Brockhoff & Christensen, 2017). Separate models were fitted for each of the dependent variables. Because the duration of the trials (and hence the total dwell time and total fixation count) was limited by the duration of the audio recordings in the RWL condition whereas reading in RO trials was self-paced, percentage measures were entered in the models as a way of controlling for differences in trial length. Following Chang and Choi's (2014) approach, the proportion of the amount of time spent gazing at texts and pictures was used, instead of the raw total reading time, for two main

reasons. First, because attention is typically used as a relative term (Cowen, 1995) and second, because percentage measures have also been used in studies on multimedia learning to study attention allocation to pictures and text (e.g., Chang & Choi, 2014; d'Ydewalle & De Bruycker 2007; Johnson & Mayer, 2012; Yang et al., 2013). In order to answer the second research question, we fitted logistic regression models to the response accuracy data using the glm function from the base R stats package.

RESULTS

Processing of text and images

The total time that learners spent processing the text and image areas in the two conditions (RO vs. RWL) was explored first (see Table 1).

Table 1. *Dwell time, fixation count, and average fixation duration descriptive statistics by condition and region (SD in brackets). Values reported are mean values per page/trial.*

Condition	Region		Dwell Time	Dwell Time %	Fixations N	Fixations %	Fixation duration
RO	Picture	M	1,258 ms	8%	6.31	10%	179 ms
		SD	(1,135)	(6.74)	(4.52)	(6.45)	(61.71)
		95% CI	1,104- 1,413	6.82-8.66	5.70-6.93	8.82-10.58	171-188
RO	Text	M	15,823 ms	92%	59.55	90%	263 ms
		SD	(5,512)	(6.79)	(15.79)	(6.4)	(48.74)
		95% CI	15,045-16,602	91.09-93.01	57.33-61.79	89.10- 90.91	257-270
RWL	Picture	M	1,861 ms	11%	8.74	13%	204 ms
		SD	(1,165)	(7.08)	(4.74)	(6.88)	(61.12)

		95% CI	1,703-2,020	10.15-12.07	8.10-9.39	12.51-14.38	196-213
RWL	Text	M	15,124 ms	89%	55.97	87%	272 ms
		SD	(3,570)	(7.1)	(11.09)	(7.02)	(50.35)
		95% CI	14,656-15,594	87.90-89.77	54.52-57.43	85.46-87.31	266-279

As explained above, percentage measures were entered in the models as a way of controlling for differences in trial length and as a measure of relative attention distribution. The three dependent measures were checked for normality using the fitdistrplus package (v 1.0-14; Delignette-Muller & Dutang, 2015) and the method outlined by Cullen and Frey (1999). They were found to significantly deviate from normality. Shapiro-Wilk tests confirmed significant deviations from normality for Dwell Time %: $W = 0.76$, $p < .0001$; Fixation Count %: $W = 0.77$, $p < .0001$; and Average Fixation Duration: $W = 0.98$, $p < .0001$. There is, however, marked disagreement as to the importance of parametric assumptions (e.g., McCulloch & Neuhaus, 2011), with evidence pointing to the relative robustness of linear mixed models to violations of normality (Arnau, Bendayan, Blanca & Bono, 2013), and debate as to the cost-benefits of data transformations to address issues of non-normality in terms of interpretability of the effects (Liceralde & Gordon, 2018). While some authors have recommended alternative statistical approaches (e.g., GLMMs; Lo & Andrews, 2015), particularly in cases of small sample sizes (Arnau et al., 2013), in this instance, we have instead opted to fit a robust linear mixed model to the data using the robustlmm package (v 2.3; Koller, 2016) as a check for our analysis. A comparison of the coefficients produced by the lme4 and robustlmm package showed broad agreement. Thus, the coefficients produced by the lme4 package are reported here.

A first model was fitted to the Dwell time % data, modelling the interaction between condition (RO vs RWL) and region of interest (TEXT vs IMAGE) as fixed effects, with random intercepts for the effect of trial and participant (a model with random slopes for the effect of trial within participants was not found to be a better fit for the data by computing an ANOVA between the two models, $\chi^2 = 0$, $p = 1$). The model (see Table S1 in the supplementary materials) revealed significant main effects of Condition, $\beta = 0.03$, t(836) = 4.97, $p < .0001$, $d = 0.34$, and of Region, $\beta = 0.84$, t(836) = 122.04, $p < .0001$, $d = 8.44$, as well as a significant interaction between the two, $\beta = -0.06$, t(836) = -6.87, $p < .0001$, $d = -0.47$. To decompose the interaction, we ran Bonferroni-corrected post-hoc comparisons between all levels of the two factors (i.e., condition and region) using the emmeans package (v 1.3.5.1). These revealed that, proportionally, the participants spent more time on the text region (compared to the picture region) in the RO condition than in the RWL condition, $\beta = -0.03$, z = -4.73, $p < .0001$, whereas more time was spent fixating the images during RWL trials than during RO trials, $\beta = 0.03$, z = -4.97, $p < .0001$.

Fixation counts on the two regions of interest were then examined (see descriptive statistics in Table 1). As explained above, these were computed as percentages of the total number of fixations recorded during each trial. The same model structures were fitted to this dependent variable as for Dwell time %. This model (see Table S2 in the supplementary materials) revealed significant main effects of both Condition, $\beta = 0.03$, t(836) = 5.71, $p < .0001$, $d = 0.39$, and Region, $\beta = 0.80$, t(836) = 120.46, $p < .0001$, $d = 8.33$, as well as a significant interaction between the two factors, $\beta = -0.07$, t(836) = -7.93, $p < .0001$, $d = -0.54$. Post-hoc comparisons on the interaction revealed the same pattern of results as the Dwell time % measure, whereby learners spent proportionally more time fixating the text region during RO trials than during RWL trials, $\beta = -0.03$, z = -5.51, $p < .0001$, whereas more time

was spent fixating the images during RWL trials than during RO trials, $\beta = 0.03$, z = 5.71, $p < .0001$.

It is worth noting that the main effects of Region, both for Dwell time % (d = 8.33) and fixation counts (d = 8.44), were considerably larger than those included in the size estimates commonly used in the literature (e.g., Cohen's levels). Effect sizes larger than 1 have been found in previous research (Hattie, 2009) and those above 2 have been described as huge in the applied statistical literature (Sawilowsky, 2009).

The average duration of fixations in the two regions of interest was then analysed (see descriptive statistics in Table 1). The model (see Table S3 in the supplementary materials) revealed a main effect of Condition, $\beta = 25.29$, t(807) = 5.31, $p < .0001$, $d = 0.37$, main effect of Region, $\beta = 83.78$, t(807) = 17.24, $p < .0001$, $d = 1.21$, and an interaction between Condition and Region, $\beta = -15.77$, t(807) = -2.33, $p < .05$, $d = -0.16$. Post-hoc comparisons showed that the difference in average fixation duration between RWL and RO trials was significant for the pictures, $\beta = 25.29$, z = 5.31, $p < .0001$, with longer average fixations on the images during trials with audio. Mean fixations on the text during RWL trials and RO trials were not significantly different, $\beta = 9.51$, z = 1.97, $p = .29$.

Text comprehension

Finally, learners' scores in the comprehension test and their relationship with viewing patterns were analysed (see response accuracy descriptive statistics in Table 2). We first looked at the relationship between processing time on a particular region (Picture or Text) and response accuracy for comprehension questions pertaining to that region type (image-only or text+image questions). Since there was not one text+image and one image-only question per page/trial, dwell times were computed per type of region of interest (Picture or Text) but averaged across trials. Similarly, participants' mean response accuracy was

computed (as a correct/incorrect ratio bounded between 0 and 1) per question type (i.e., text+image questions and image-only questions). Because the data was averaged across trials, linear mixed models were no longer viable as a statistical method; furthermore, because of the bounded nature of the outcome variable, a simple linear regression would similarly not be indicated. We therefore opted to fit logistic regression models to the response accuracy data using the glm function from the base R stats package.

Table 2. *Descriptive statistics for comprehension scores by condition and type of question (image-related or text-related) (SD in brackets).*

Condition	Image % Correct	Text % Correct
RO	44.6% (4.7)	53.57% (5.0)
RWL	42.5% (4.0)	58.48% (5.3)

Results of the logistic regression models reported in Table 3 revealed only a main effect of Region, suggesting overall better response accuracy for the text+image questions compared to image-only questions. There was no main effect of Condition, suggesting a similar response accuracy in the RWL and RO conditions. There was also a significant interaction between Dwell time % and Region, with a higher % of fixations on the text related to lower accuracy. The interaction between Condition and Region was also significant, suggesting that there is a stronger relationship between Dwell time % and comprehension accuracy in the RO compared to the RWL condition (see Figure 1).

Table 3. *Coefficients of the logistic regression model fitting average Dwell time % on the relevant region type (text, picture) against average response accuracy for questions pertaining to that region type.*

	Average Response Accuracy					
Predictors	*β*	*B*	*β CI 95%*	*z*	*p*	*OR*
(Intercept)	-0.75	-	-1.42 – -0.10	**-2.24**	**.02**	
AvgDT (%)	7.17	11.37	0.44 – 14.41	**2.03**	**.04**	**1308**
ConditionRWL	0.25	0.50	-0.95 – 1.46	0.41	.67	1.29
RegionTEXT	5.74	11.15	1.01 – 10.74	**2.32**	**.02**	**311.28**
AvgDT:ConditionRWL	-5.45	-7.88	-16.59 – 5.52	-0.97	.33	.004
AvgDT:RegionTEXT	-12.43	-21.92	-21.33 – -3.96	**-2.82**	**.004**	**<.0001**
ConditionRWL:RegionTEXT	8.60	14.47	0.54 – 16.86	**2.07**	**.03**	**5434**
AvgDT:ConditionRWL:RegionText	-4.45	-6.66	-18.64 – 9.73	-061	.53	.01

Intercept for *Condition* = -0.49
Intercept for *Region* = 4.99

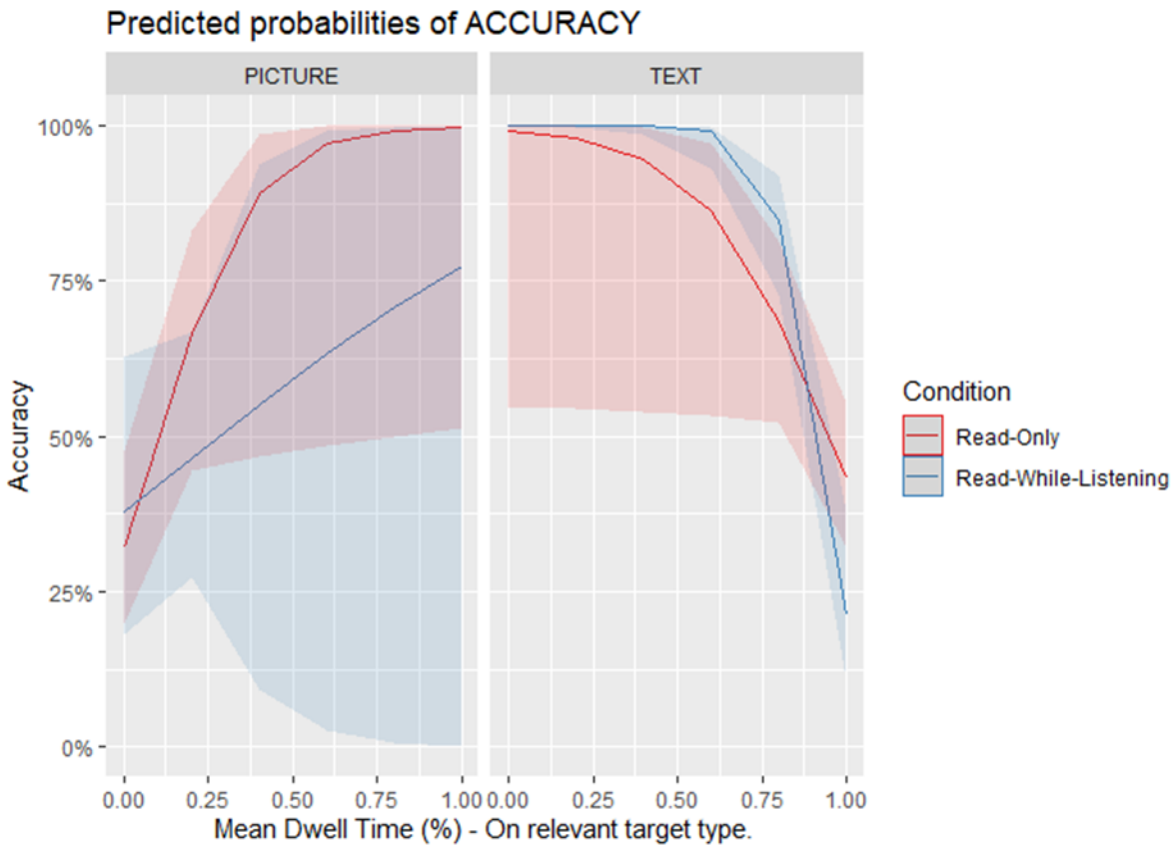


Figure 1. Predicted probability of comprehension accuracy on text+image and image-related questions as a function of Dwell time % on the relevant target type in RO and RWL conditions.

Because the three-way interaction was not significant, a final model was fitted with only the two-way interactions between Dwell time % and Region, and between Condition and Region (see results in Table 4). In line with the findings of the previous model, this model revealed a main effect of Region and a significant interaction between Dwell time % and Region. The interaction between Condition and Region was not significant in this model, suggesting that the relationship between Dwell time % and comprehension accuracy does not differ by condition.

Table 4. *Coefficients of the logistic regression model fitting average Dwell time % on the relevant region type (text, picture) against average response accuracy for questions pertaining to that region type (two-way interactions only).*

	Average Response Accuracy					
Predictors	*β*	*B*	*β CI 95%*	*z*	*p*	*OR*
(Intercept)	-0.58	-	-1.14 – -0.03	**-2.06**	**.03**	**0.55**
AvgDT (%)	5.07	8.05	-0.17 – 10.49	1.87	.06	160.25
ConditionRWL	-0.28	-0.54	-0.81 – 0.24	-1.04	.29	0.75
RegionTEXT	9.00	17.50	5.09 – 13.12	**4.40**	**<.0001**	**8152**
AvgDT:RegionTEXT	-14.05	-24.77	-20.97 – -7.31	**-4.04**	**<.0001**	**<.0001**
ConditionRWL:RegionTEXT	0.18	0.31	-0.48 – 0.85	.54	.58	1.20

Intercept for *Condition* = -0.86
Intercept for *Region* = 8.42

This analysis shows the connection between percentage of time on a particular type of region and accuracy in responding to questions that referred to that area (i.e. percentage of time on images and accuracy on image-only questions, as well as percentage of time on the text and accuracy on the text+image questions). However, it could also be hypothesized that processing time on a type of region (Picture or Text) might also support comprehension of questions related to the opposite target type. The percentage of time on images might support comprehension of text+image questions, and percentage of time on text could also support comprehension of image-only questions. Thus, as a final analysis we tested whether Dwell time % on the opposite target type was related to response accuracy (i.e., whether percentage of time spent looking at the picture could help in correctly answering text+image questions, and vice versa). To do this, we swapped the eye movement measures between regions of interest and used the new resulting variable as a predictor in the same logistic regression structure used in the previous analyses. The results showed that there was a significant

interaction between Dwell time % and Region, β = 12.59, z = 2.85, p = .004, suggesting that spending proportionally more time looking at the images was related to greater accuracy on text+image questions, but proportionally more time spent looking at the text was related to lower accuracy on image-only questions (see Figure 2).

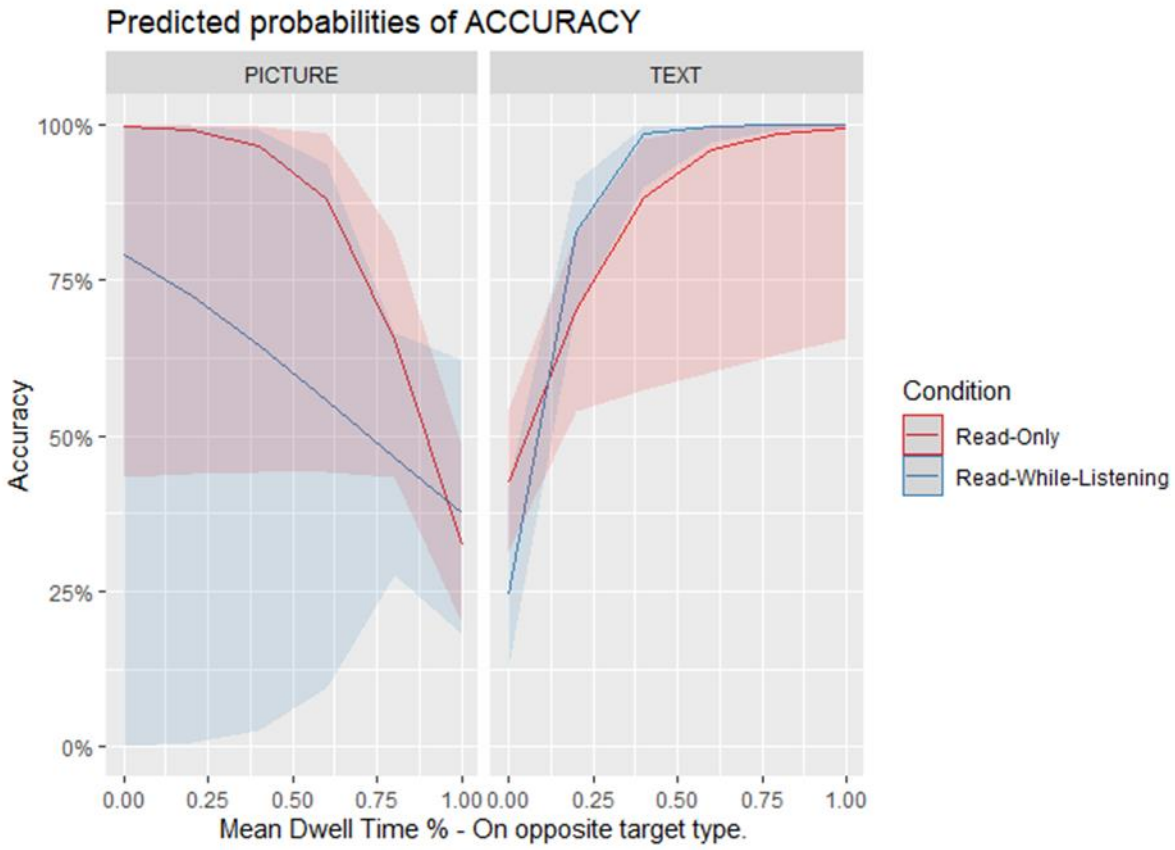


Figure 2. Predicted probability of comprehension accuracy on text+image and image-related questions as a function of Dwell time % on the opposite target type in RO and RWL conditions.

It must be pointed out that the logistic regressions reported here produced odds ratios that were unusually large and, in a couple of cases, unusually small. We report them in the table because, for significant effects, they nevertheless had confidence intervals that – however wide – reliably did not cross 0. This is also consistent with the wide CIs observed in Figure 1 and Figure 2, and likely the result of a low number of observed events for each accuracy level.

DISCUSSION

The current study contributes important knowledge about the processing of multimodal materials in L2 and EFL contexts. Importantly, it responds to the call for more research on the relationship between eye-movement patterns and learning outcomes (Alemdag & Cagiltay, 2018). To achieve this, we examined young learners' processing of images and written text in the presence and absence of an auditory version of the text. More specifically, we looked at the proportion of fixations and proportion of fixation duration on the text and image areas. We also explored the relationship between reading/viewing patterns and performance on a comprehension test. Results of the analysis of Dwell time % and Fixation count % showed that in general, young L2 learners are likely to spend proportionally more time on the processing of the text than the images in both multimodal conditions (RO and RWL). This confirms previous research showing that, when presented with pictures and text, learners tend to spend more time on the text (Schmidt-Weigand, Kohrert, & Glowalla, 2010). It is important to note that these patterns could be explained by the degree of informativeness of the different input sources in the materials. In this study, the text carried most of the information and learners were also aware that they would be answering some comprehension questions after the reading task. Thus, it is not surprising that they spent more time processing the text. Different patterns would be expected with other types of reading materials where the visual input carries most of the information, such as in comic books.

In response to the first research question, results showed that processing patterns were clearly affected by the presentation of the auditory input. The analyses demonstrated that in the RWL mode young learners spend proportionally more time and have more fixations on the pictures than in RO conditions, while in the RO mode they spend proportionally more time and have a higher percentage of fixations on the text than they do in the RWL mode. Average fixation durations were also longer in general in the RWL condition, particularly for the images. As Serrano and Pellicer-Sánchez (2019) argue, because the verbal input is presented auditorily, learners can look at the pictures more often and make a better use of them, which allows them to better integrate the verbal and non-verbal sources of input. Looking more at the pictures in the RWL condition does not seem to hinder comprehension. The lack of differences in comprehension between the RO and RWL conditions does not support an advantage of the RWL over RO as it was the case in earlier investigations (e.g., Chang, 2009; Chang & Millet, 2015), nor a detrimental effect of RWL on comprehension (e.g., Diao & Sweller, 2007). Importantly, it supports results of previous investigations showing the beneficial effect of RWL for young learners' comprehension (e.g., Lightbown, 1992).

The current findings go against the redundancy principle, which suggests a negative effect of presenting a text in both written and spoken modalities. Kalyuga and Sweller (2014) argued that this negative effect should be particularly evident for L2 readers because of the difficulty they have linking auditory and written input. However, results of the present study show similar levels of comprehension in the RO and RWL conditions. This suggests that the negative effect of the presentation of written and spoken input found in the context of L1 content learning (e.g., Kalyuga, Chandler, & Sweller, 1999; Jamet & Le Bohec, 2007; Mayer, Heiser, & Lonn, 2001) is not applicable to the L2 reading context, at least with young learners. Our results are also problematic for the dual modality principle, which predicts a

detrimental effect of dual modality presentation. The results of the present study show processing differences when the verbal input is provided through two modalities (i.e., spoken and written), but this does not seem to have a detrimental effect on comprehension, supporting previous research findings (e.g., Chang & Millet, 2015). Again, this calls into question the applicability of principles of multimedia learning in L2 contexts.

In addition, the proportionally longer reading times for the text in the RO condition could be a reflection of the possibility of pausing and re-reading parts of the passage; this would allow readers to adapt their processing speed to their needs, as suggested by Schnotz (2014). Results of the present study also confirm the findings of the study by Serrano and Pellicer-Sánchez (2019), which was also conducted with young learners of similar age and proficiency, and demonstrate similar reading and viewing patterns with their unmodified, more authentic materials and our less authentic, more controlled ones.

Concerning the eye movements examined in the present study, the analyses of the average fixation durations have shown that, contrary to what was expected, fixations on images were shorter than fixations on the text. This is in contrast to what has been suggested for adult readers and might be a reflection of the developing reading skills of the participants in the present study. As expected, average fixation durations on the text were longer than typical reading times in adult readers (Rayner, 1998, 2009; Whitford & Joanisse, 2018).

In response to the second research question, results of this study have shown that a higher percentage of total dwell time on the text was related to lower accuracy on the text+image questions. This is in line with previous findings suggesting that longer relative gaze duration (i.e., proportion of the amount of time gazing at text) (e.g., Chang & Choi, 2014) and longer total dwell time (e.g., Serrano & Pellicer-Sánchez, 2019) are a sign of processing difficulties that are then reflected in lower comprehension scores. Importantly,

results of the present study have provided initial evidence of a relationship between the amount of attention allocated to images and scores on text+image questions. A higher proportion of processing time on images supported comprehension of the text, providing evidence for the positive role of images in reading comprehension. This finding is also in line with previous studies showing a positive relation between total dwell time (Bisson, et al., 2015) as well as relative attention (i.e., fixation count on picture – fixation count on text) (Eitel, 2016) on pictures and learning scores. It could also be hypothesised that the relationship between the processing of the images and accuracy on text+images related questions is a consequence of faster readers spending proportionally less time on the text. Spending less time on the text, and consequently having more time available to look at the images, could be a sign of higher proficiency in reading that is then reflected in accuracy scores. Interestingly, more time on images was not related to higher accuracy on the image-only questions, suggesting that when processing the images, learners did not pay particular attention to the specific visual features that the questions addressed and that they used images mainly as support for text comprehension.

The results of this study have important implications for teaching. The present study indicates that while the children spent more time on the images in the RWL mode, their comprehension was equally good in both modes. We know from previous studies that RWL is popular with young learners and that they generally show positive attitudes towards RWL (e.g., Lightbown et al., 2002; Tragant, Muñoz, & Spada, 2016; Tragant & Vallbona, 2018). Based on this evidence it seems advisable for teachers to promote RWL among young learners. It is a powerful language learning tool in less formal contexts both at school and at home. Importantly, the present study also supports the use of images to support young learners' comprehension. Despite claims that images may pull attention away from the text

(e.g., Hill, 2013), this study has shown that the amount of attention allocated to images seems to support reading comprehension.

It is important to acknowledge the limitations of the present study. This investigation is the first to examine the potential role that the percentage of dwell time allocated to images has on comprehension. However, the number of images used in the present study, as well as the content they depicted, did not allow us to have a larger number of image-only questions. Having enough questions that can only be answered by processing the images might only be possible in a much longer experiment and/or with images drawn specifically for the purposes of the study. Finally, the results of the present study shed light on our understanding of young L2 learners' processing of multimodal input, but it remains to be demonstrated whether similar patterns would be observed with learners of different ages and proficiency levels. Future studies should examine the relationship between reading and viewing patterns and comprehension with L2 learners having a wide variety of characteristics.

CONCLUSION

This study has provided further evidence for the benefits of using eye-tracking to examine processing during multimodal learning (Mayer, 2017). The results of the present study show that the addition of auditory input leads to processing differences, with proportionally more time spent on images in the RWL condition than in RO. These processing differences suggest a better integration of the verbal and pictorial sources of information in multimedia materials with auditory input, without having a negative impact on comprehension. Importantly, this study has shown that proportionally more time on text is related to lower levels of comprehension, whereas more time on images is related to better comprehension, revealing interesting differences in the relationship between reading/viewing patterns and comprehension.

NOTES

[1] While it is likely that participants with a vocabulary size of 1,985 words would know most of the words in the 1K, a VST does not provide information of word knowledge at different frequency levels. Future studies should use a vocabulary measure that provides more reliable information about knowledge at different frequency levels.

[2] Normal speech rate in English is approximately 150 wpm (Buck, 2001; Chang, 2011; Griffiths, 1990). This is the speech rate followed in audiobooks for adults. Examination of the audio recordings of graded readers at this level showed that the speech rate ranged from 90-120 wpm.

ACKNOWLEDGEMENTS

This manuscript is an output from an ELT Research Award funded by the British Council to promote innovation in English language teaching research. The views expressed are not necessarily those of the British Council. We would like to thank Fabio Parente and Ling Shuping (University of Nottingham), and Radha Chandy (University of Barcelona) for their help with data collection and analysis.

REFERENCES

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education, 125*, 413–428.

Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods, 45*(3), 873-879.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bates D., Maechler M., Bolker B., & Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Bisson, M-J., Van Heuven, W., Conklin K., & Tunney, R. (2014). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics, 35,* 399-418. doi: 10.1017/S0142716412000434

Bisson, M-J., Van Heuven, W., Conklin, K., & Tunney R. (2015). The role of verbal and pictorial information in multi-modal incidental acquisition of foreign language vocabulary. *Quarterly Journal of Experimental Psychology, 68,* 306–26. doi: 10.1080/17470218.2014.979211

Brown, R., Waring, R., & Donkaewbua S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*(2), 136-163.

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

Chang, C.-S. (2009). Gains to L2 listeners from reading while listening vs. listening only in comprehending short stories. *System, 37,* 652-663. doi: 10.1016/j.system.2009.09.009

Chang, C.-S. (2011). The effect of reading while listening to audiobooks: Listening fluency and vocabulary gain. *Asian Journal of English Language Teaching, 21*, 43–64.

Chang, Y., & Choi, S. (2014). Effects of seductive details evidenced by gaze duration. *Neurobiology of Learning and Memory, 109,* 131–138.

Chang, C.-S., & Millett, S. (2014). The effect of extensive listening on developing l2 listening fluency: Some hard evidence. *ELT Journal, 68*(1), 31-40. doi: 10.1093/elt/cct052

Chang, C.-S., & S. Millett, S. (2015). Improving reading rates and comprehension through audio-assisted extensive reading for beginner learners. *System, 52*, 91-102. doi: 10.1016/j.system.2015.05.003

Cobb, T (n.d.). *Lextutor. Vocabprofile* [computer program]. Accessed at https://www.lextutor.ca/cgi-bin/range/texts/index.pl

Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for Applied Linguistics Research*. Cambridge: Cambridge University Press.

Cullen, A. C., & Frey, H. C. (1999). *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs.* New York: Plenum Press.

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software, 64*(4), 1-34.

Diao, Y., & Sweller, J. (2007). Redundancy in foreign language reading comprehension instruction: Concurrent written and spoken presentations. *Learning and Instruction, 17*, 78-88.

D'Ydewalle, G., & de Bruycker W. (2007). Eye movements of children and adults while reading television subtitles. *European Psychologist, 12*(3), 196-205. doi: 10.1027/1016-9040.12.3.196

Eitel, A. (2016). How repeated studying and testing affects multimedia learning: Evidence for adaptation to task demands. *Learning and Instruction, 41*, 70–84. doi:10.1016/j.learninstruc.2015.10.003.

Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sarkar, A. & Yoon, H. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition, 21*(3), 563-584. doi:10.1017/S1366728917000219.

Griffiths, R. (1990). Speech rate and nonnative speaker comprehension: A preliminary study in the time-benefit analysis. *Language Learning, 40*, 311–336.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement.* Oxon: Routledge.

Hill, D. (2013). Graded readers. *ELT Journal, 67*(1), 85-125.

Hochpöchler, U., Schnotz, W., Rasch, T., Ullrich, M., Horz, H., McElvany, H., & Baumert, J. (2013). Dynamics of mental model construction from text and graphics. *European Journal of Psychology of Education, 28*(4), 1105–1126. doi: 10.1007/s10212-012-0156-z

Hu, M., & Nation, I.S.P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Jamet, E. & Le Bohec, O. (2007). The effect of redundant text in multimedia instruction. *Contemporary Educational Psychology, 32*, 588-598.

Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied, 18*(2), 178–191. doi: 10.1037/a0026923

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13,* 351-371.

Kalyuga, S. & Sweller, J. (2014). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed.) (pp. 247-262). New York, NY: Cambridge University Press.

Koller, M. (2016). robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. *Journal of Statistical Software, 75*(6), 1-24. doi:10.18637/jss.v075.i06

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1–26. doi: 10.18637/jss.v082.i13

Liceralde, V. R. T., & Gordon, P. C. (2018, October 23). Consequences of power transforms as a statistical solution in linear mixed-effects models of chronometric data. Retrieved from osf.io/ygc7s

Lightbown, P.M. (1992). Can they do it themselves? A comprehension-based ESL course for young children. In R. Courchene, J. St John, C. Therien, & J. I. Glidden (Eds.), *Comprehension-based Second Language Teaching* (pp. 353–370). Ottawa: University of Ottawa Press.

Lightbown, P., Halter, R., White, J., & Horst, M. (2002). Comprehension-Based Learning: The Limits of 'Do It Yourself'. The Canadian Modern Language Review, 58(3), 427-464. doi:10.3138/cmlr.58.3.427

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6,* 1171. doi:10.3389/fpsyg.2015.01171

Mason, L., Pluchino, P., Tornatora, M. C., & Ariasi N. (2013). An eye-tracking study of learning science text with concrete and abstract illustrations. *Journal of Experimental Education, 81*(3), 356-384. doi: 10.1080/00220973.2012.727885

Mason, L., Tornatora M. C., & Pluchino, P. (2015). Integrative processing of verbal and graphical information during re-reading predicts learning from illustrated text: An eye movement study. *Reading and Writing, 28,* 851-872. doi: 10.1007/s11145-015-9552-5

Massaro D.W. (2012). Multimodal Learning. In N. M. Seel (Eds.) *Encyclopedia of the sciences of learning* (pp. 2375-2378). Boston, MA: Springer.

Mayer, R. E. (2001). *Multimedia learning*. New York, NY: Cambridge University Press.

Mayer, R. E. (2009). *Multimedia Learning* (2nd ed.). New York, NY: Cambridge University Press.

Mayer, R. E. (2014a). Introduction to multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed.) (pp. 1-24). New York, NY: Cambridge University Press.

Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning*, 1–21. doi: 10.1111/jcal.12197

Mayer, R., Heiser, J., & Lonn (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93,* 187-198.

Mayer R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology. 90,* 312-320.

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words?: Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology, 86*(3), 389-401.

Meara, P.M., & Milton, J.L. (2003). *X_Lex: The Swansea Vocabulary Levels Test*. Newbury: Express Publishing.

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science,* 26(3), 388-402. doi: 10.1214/11-STS361

Montero Perez, M., Peters, E., Clarebout, G., Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning and Technology, 18(*1), 118-141.

Montero Perez, M., Peters, E., & Desmet P. (2015). Enhancing vocabulary learning through captioned video: An eye-tracking study. *The Modern Language Journal, 99,* 308–28. doi: 10.1111/modl.12215

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91,* 358-368.

Moussa-Inaty, J., Ayres, P., & Sweller, P. (2012). Cognitive load and the impact of spoken English on learning English as a foreign language. *Applied Cognitive Psychology, 63,* 391-402.

Muñoz, C. (2017). The role of age and proficiency in subtitle reading. An eye-tracking study. *System, 67*, 77-86. doi:10.1016/j.system.2017.04.015

Niegeman, H., & Heidig, (2012). Multimedia Learning. In N. M. Seel (Eds.) *Encyclopedia of the sciences of learning* (pp. 2372-2375). Boston, MA: Springer.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.

Paivio, A. (2006). *Mind and its evolution: A dual coding approach*. Mahwah, NJ: Lawrence Erlbaum.

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition, 38*, 97– 130. doi: 10.1017/S0272263115000224.

Pellicer-Sánchez, A., & Conklin, K. (2020). Eye-tracking as a data collection method. In H. Rose and J. McKinley (Eds*.), The Routledge Handbook of Research Methods in Applied Linguistics.* New York, NY: Routledge.

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly, 53*(4), 1008-1032. doi: 10.1002/tesq.531

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*, 1457–1506.

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and Instruction, 20*(2), 100–110. doi: 10.1016/j.learninstruc.2009.02.011

Schnotz, W. (2014). An integrated model of text- and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 72-103). New York, NY: Cambridge University Press.

Serrano, R., & Pellicer-Sánchez, A. (2019). Young L2 learners' online processing of information in a graded reader during reading-only and reading-while-listening conditions: A study of eye movements. *Applied Linguistics Review*, First view. doi: 10.1515/applirev-2018-0102

Sung, E., & Mayer, R. E. (2013). Online multimedia learning with mobile devices and desktop computers: An experimental test of Clark's methods-not-media hypothesis. *Computers in Human Behavior, 29,* 639-647. doi: 10.1016/j.chb.2012.10.022

Sawilowsky, S. S. (2009) New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods, 8* (2), 597-599. doi: 10.22237/jmasm/1257035100

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257-285. doi: 10.1207/s15516709cog1202_4

Taguchi, E., Takayasu-Maass, M., & Gorsuch G. (2004). Developing reading fluency in EFL: How assisted repeated reading and extensive reading affect fluency development. *Reading in a Foreign Language, 16*(2), 70-96.

Tragant, E., & Pellicer-Sánchez. A. (2019). Young learners' engagement with multimodal exposure: An eye-tracking study. *System, 80*, 212-223. doi: 10.1016/j.system.2018.12.002

Tragant, E. & Vallbona, A. (2018). Reading while listening to learn: Young EFL learners' perceptions. *ELT Journal, 72*(4), 395-404. doi: 10.1093/elt/ccy009

Trofimovich, P., Lightbown, P.M., Halter, R., & Song, H. (2009). Comprehension based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition, 31,* 609–639.

Tragant, E., Muñoz, C., & Spada, N. (2016). Maximizing young learners' input: An intervention program. *The Canadian Modern Language Review, 72,* 234–257.

Warren, P., Boers, F., Grimshaw, G., & Siyanova-Chanturia, A. (2018). The effect of gloss type on learners' intake of new words during reading. *Studies in Second Language Acquisition, 40*(4), 883-906. doi:10.1017/S0272263118000177

Webb, S., & Chang A. (2012). Vocabulary learning through assisted and unassisted repeated reading. *The Canadian Modern Language Review*, *68*(3), 267-290. doi: 10.3138/cmlr.1204.1.

Webb, S., & Chang, A. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research, 19,* 667-686.

Webb, S., & Chang, A. C-S. (2017). How Does Mode of Input Affect the Incidental Learning of Multiword Combinations? Paper presented at the AAAL Annual conference, Portland, Oregon (USA).

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning, 63*(1), 91-120.

Whitford, W., & Joanisse, M. F. (2018). Do eye movements reveal differences between monolingual and bilingual children's first-language and second-language reading? A focus on word frequency effects. *Journal of Experimental Child Psychology, 173,* 318–337.

Wilde, O. (2012). *The Canterville Ghost.* Retold by Cadwallader, J. ELI Graded Readers. Recanati: ELI Publishing.

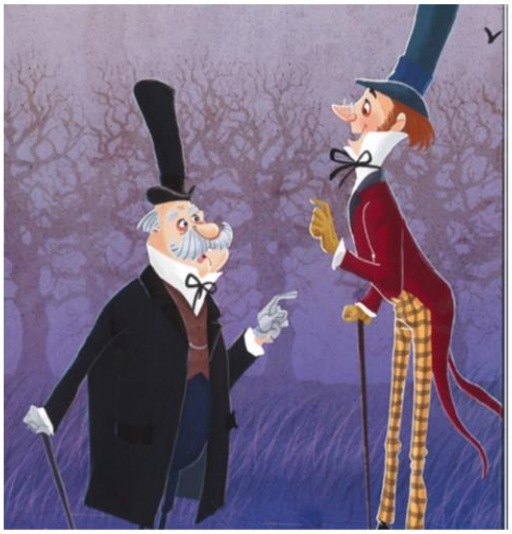
Winke, P., Gass, S., & Sydorenko T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal, 97,* 254–75. doi: 10.1111/j.1540-4781.2013.01432.x

Wright, A. (2010). *Pictures for language learning.* Cambridge: Cambridge University Press.

Yang, F., Chang, C., Chien, W., Chien, Y., & Tseng, T. (2013). Tracking learners' visual attention during a multimedia presentation in a real classroom. *Computers and Education, 62,* 208-220. doi: 10.1016/j.compedu.2012.10.009

APPENDIX

Mr Otis was a businessman from America. He wanted to buy Lord Canterville's house but Lord Canterville said, "I have to tell you Mr Otis, there is a ghost in this house". Mr Otis did not believe in ghosts so he bought the house.



Sample experimental stimuli. Text and pictures modified from The Canterville Ghost (Wilde, 2012), by ELI Publishing (https://www.elionline.com/eng/graded-readers/)

ON-LINE SUPPLEMENTARY MATERIALS

Table S1. *Coefficients of the linear mixed-effect model fitting Dwell time % against condition (RWL vs RO) and region (Text vs Picture).*

	Dwell Time %					
Predictors	*β*	*B*	*β CI 95%*	*t*	*p*	*d*
(Intercept)	.077	-	.06 – .08	**16.14**	**<.0001**	**-**
ConditionRWL	.033	.041	.02 – .04	**4.97**	**<.0001**	**0.34**
RegionTEXT	.841	1.02	.82 – .85	**122**	**<.0001**	**8.44**
ConditionRWL:RegionTEXT	-.065	-.071	-.08 – -.04	**-6.87**	**<.0001**	**-0.47**

Table S2. *Coefficients of the linear mixed-effect model fitting Fixation % against condition (RWL vs RO) and region (Text vs Picture).*

	Fixation %					
Predictors	*β*	*B*	*β CI 95%*	*t*	*p*	*d*
(Intercept)	.097	-	.08 – .10	**20.97**	**<.0001**	**-**
ConditionRWL	.037	.048	.02 – .05	**5.71**	**<.0001**	**0.39**
RegionTEXT	.802	1.03	.78 – .81	**120.46**	**<.0001**	**8.33**
ConditionRWL:RegionTEXT	-.073	-.083	-.09 – -.05	**-7.93**	**<.0001**	**-0.54**

Table S3. *Coefficients of the linear mixed-effect model fitting average Fixation Duration against condition (RWL vs RO) and region (Text vs Picture).*

	Fixation Duration					
Predictors	*β*	*B*	*β CI 95%*	*t*	*p*	*d*
(Intercept)	179.37	-	167.34 – 191.39	**29.40**	**<.0001**	**-**
ConditionRWL	25.29	.185	15.96 – 34.61	**5.31**	**<.0001**	**0.37**
RegionTEXT	83.78	.614	74.26 – 93.30	**17.24**	**<.0001**	**1.21**
ConditionRWL:RegionTEXT	-15.77	-.102	-29.02 – -2.53	**-2.33**	**.01**	**-0.16**