The Effect of Auditory Input on Multimodal Reading Comprehension: An Examination of Adult Readers' Eye Movements

Ana Pellicer-Sánchez[1], Kathy Conklin[2], Michael P. H. Rodgers[3], Fabio Parente[4]

[1] UCL Institute of Education, [2] University of Nottingham, [3] Carleton University, [4] University of Derby

ABSTRACT

Comprehension of many types of texts involves constructing meaning from text and pictures. However, research examining how L2 learners process text and pictures and its relationship with comprehension is scarce. Thus, while verbal input is often presented in written and auditory modes simultaneously (i.e., audio of text with simultaneous reading of it), we do not know how the auditory input affects L2 adult learners' processing of text and pictures and its relation to comprehension. In the current study, L2 adult learners and L1 adults read and read-while-listening to an illustrated story while their eye movements were recorded. Immediately after reading, they completed a comprehension test. Results showed that the presence of auditory input allowed learners to spend more time looking at pictures and supported a better integration of text and pictures. No differences were observed between L2 and L1 readers' allocation of attention to text and pictures. Both reading conditions led to similar levels of comprehension. Processing time on the text was positively related to comprehension for L2 speakers, while it was associated to lower comprehension for L1 speakers. Processing time on images was positively related to comprehension only for L1 speakers.

Keywords: multimodal reading; multimedia learning; second language; reading-while-listening; eye-tracking; reading comprehension

INTRODUCTION

Multimodality is gaining attention in second language (L2) learning and teaching. Multimodal information is presented in a combination of modes, i.e., language, image, gesture, movement, music or sound. Current L2 reading practices are largely multimodal, in both printed text and digital media, with learners being regularly exposed to a combination of written text, images, and often auditory input. Successful comprehension of multimodal texts involves the ability to allocate attention to text and pictures and integrate them. However, our understanding of how L2 learners engage with multimodal texts and the relationship with comprehension is still limited. A few empirical studies on L2 multimodal reading have been conducted to date showing that the combination of images and text in multimodal reading supports comprehension (e.g., Omaggio, 1979; Plass, Chun, Mayer & Leutner, 1998) and vocabulary learning (e.g., Plass et al., 1998).

In the last decade, there has been an increase in the number of eye-tracking studies looking at readers' interaction with multimodal texts. The use of eye-tracking allows researchers to obtain a direct indication of cognitive processing during multimedia learning (Mayer, 2017). In the first language (L1), research has shown that, when exposed to pictures and printed text, readers spend most of the time processing the text, inspecting pictures only minimally (e.g., Johnson & Mayer, 2012; Schmidt-Weigand, 2011; Schmidt-Weigand, Kohnert, & Glowalla, 2010). When the verbal input is presented auditorily (as opposed to visually, i.e., written), more attention is allocated to the images (e.g., Schmidt-Weigand, Kohnert, & Glowalla, 2010). The integration of information presented in text and pictures seems to be key for comprehension in multimodal reading and eye-tracking studies have shown that this integration supports learning (e.g., Mason, Tornatora, & Pluchino, 2015).

However, research has mainly been conducted in the context of domain learning in the L1, primarily in relation to learning from science texts (Alemdag & Cagiltay, 2018). Few eye-tracking studies on multimodal reading have been conducted in the L2 learning context. In addition, despite the reported benefits of the simultaneous presentation of written and auditory input for comprehension in the L2 (e.g., Chang & Millet, 2014, 2015), very little is known about how L2 learners process and integrate text and pictures when auditory input is also provided (i.e., the text is read aloud). Recent eye-tracking studies provided initial evidence that the presence of auditory input in multimodal reading materials containing text and pictures leads to processing differences, with more attention to the images in the presence of auditory input (e.g., Pellicer-Sánchez, et al., 2020; Serrano & Pellicer-Sánchez, 2019). This suggests that the addition of auditory input in multimodal reading materials might facilitate integration of text and images. These recent investigations have also shown that processing time on the text seems to be negatively related to comprehension (e.g., Serrano & Pellicer-Sánchez, 2019), whereas processing time on images seems to have a positive effect on comprehension (Pellicer-Sánchez, et al., 2020).

Crucially, these studies were conducted with young EFL learners (11-12 years old). Thus, the processing patterns found could largely be explained by young learners' developing reading and multimodal literacy skills. While L2 adult readers regularly engage in multimodal reading as well, very little is known about how they process different types of multimodal materials and the relationship between processing patterns and comprehension. Furthermore, previous studies have not examined how L2 processing patterns in multimodal reading compare to those of adult L1 readers. The present study addressed these gaps by examining adult L2 learners' allocation of attention to text and images in multimodal materials in the presence and absence of auditory input (i.e., the text being read aloud), as well as its impact on comprehension. In addition, L2 processing patterns were compared to L1 patterns, in an attempt to build a comprehensive picture of readers' interaction with multimodal materials.

BACKGROUND

Multimodal and multimedia second language reading

The term multimodality was coined by Kress and Van Leeuwen (1996). Multimodality refers to the construction of meaning through a variety of modes, such as language, image, gesture, movement, music or sound (Jewitt & Kress, 2003). Multimodal reading involves the combination of written text with other modes, usually static images and/or audio. Comprehension of many texts involves the integrative construction of meaning from both text and images (Daly & Unsworth, 2011; Unsworth, Thomas & Bush, 2004). The ability to integrate the text and images in multimodal texts is considered an important aspect of reading comprehension and the pictorial-verbal interface has been regarded as a crucial dimension of literacy development (e.g., Bezemer & Kress, 2008; Kress, 2000; Unsworth, 2014). As Unsworth (2014) argues, "it is now inadequate to consider reading simply as processing information in print" (p. 26).

The majority of studies on multimodal reading in the L2 context build on Mayer's influential Cognitive Theory of Multimedia Learning (Mayer, 2001, 2009, 2014a). This theory aims at understanding how learners construct meaning from words and pictures. According to the theory, when processing pictorial and verbal information simultaneously, learners engage in five processes: a) selecting relevant words from the presented text (written or aural); b) selecting relevant images from the presented graphics; c) organizing the selected words into a coherent verbal representation; d) organizing selected images into a coherent pictorial representation; e) and integrating the pictorial and verbal representations and prior knowledge (Mayer, 2014b). As Mayer (2014b) argues, the most crucial step in multimedia learning is the integration of representations from words and pictures.

Multimedia environments are presented in different modes, i.e., pictorial versus verbal. This presentation modality is related to Paivio's Dual Coding Theory (1986, 2006), which explained the relationship between verbal and non-verbal systems and suggested that the simultaneous activation of verbal and non-verbal systems supports learning. Apart from the different presentation modes (i.e., pictorial and verbal), multimedia materials can also be presented through different sensory modalities i.e., visual, auditory, and kinaesthetic.

Based on empirical evidence, Mayer (2009) identified 12 main principles for successful multimedia learning. Three of these principles are particularly relevant for the type of multimodal L2 reading investigated in the present study: *multimedia, redundancy,* and *modality.* The main tenet of Mayer's theory (i.e., the *multimedia principle*) is that people learn better from words (spoken or written) and pictures (e.g., illustrations, photos, animations and video) than from words alone (Mayer, 2001, 2009, 2014a). According to the *redundancy principle*, "people learn better from graphics and narration than from graphics, narration and printed text" (Niegeman & Heidig, 2012, p. 2374). In other words, presenting the same information in different forms (e.g., text in printed and auditory forms) is considered redundant and could interfere with learning (Kalyuga & Sweller, 2014). More specifically, it is thought that redundant information can overload working memory, having a negative impact on comprehension and learning. Finally, the *modality principle* posits that presenting information in the same mode leads to split attention, which can have a detrimental effect on learning. Thus, presenting pictures with auditory text would be more beneficial than presenting them with written text (Mayer & Moreno, 1998; Moreno & Mayer, 1999), as the simultaneous presentation of written text and pictures leads to split attention and could have a detrimental effect on learning (Schnotz, 2014).

Mayer's Cognitive Theory of Multimedia Learning and its principles have been used to support the use of images (both static and dynamic) in L2 language learning. However, most of the empirical evidence supporting it comes from domain learning in the L1 (e.g., learning mathematics in the L1). Crucially, scholars have emphasized the need to distinguish between domain learning and language learning, as they have very different learning goals (Abraham & Farias, 2017; Schnotz & Baadte, 2008). As Plass and Jones (2005) argue, "some of the other multimedia learning principles that were found in the context of science learning are not likely to extend to the area of second-language acquisition without modification" (p. 480). For example, in relation to the redundancy and modality principles, Plass and Jones (2005) highlighted the positive role that the simultaneous presentation of auditory and written text could have for L2 learners, as one can be used as input enhancement for the other. Indeed, previous studies have provided evidence against the redundancy and modality principles for L2 vocabulary learning (e.g., Farias et al., 2014) and comprehension (e.g., Pellicer-Sánchez et al., 2020; Serrano & Pellicer-Sánchez, 2019).

Despite the claims that images accompanying texts (e.g., graphs, charts, photographs, and illustrations) need to be better exploited in the L2 classroom (Abraham & Farias, 2017; Royce, 2002), few empirical studies have explored combining texts and pictures. The few existing studies have shown that images in multimodal reading led to improved comprehension (e.g., Omaggio, 1979; Plass et al., 1998) and to increased vocabulary learning (e.g., Plass et al., 1998). However, more research is needed to gain a better understanding of how L2 learners engage with multimodal texts and examine the potential relationship between engagement with the input and comprehension. In addition, it is important to explore the applicability of the principles of multimedia learning to L2 reading.

L2 reading-while-listening

Reading-while-listening, or assisted reading, involves the simultaneous presentation of the same text in written and aural formats. Numerous studies have explored the potential benefits of reading-while-listening for L2 learning. Research conducted with adult learners has shown that reading-while-listening supports vocabulary learning (e.g., Brown, Waring, & Donkaewbua, 2008; Chang, 2011; Han & Chen, 2010; Webb & Chang, 2012, 2015), with some studies suggesting an advantage of reading-while-listening over reading-only (e.g., Webb & Chang, 2012) and listening-only (e.g., Brown, Waring, & Donkaewbua, 2008). The combination of written and auditory input has also been reported to support comprehension. Chang and Millet (2015) compared the effects of reading-while-listening versus reading-only on the reading rate and comprehension of secondary English as foreign language (EFL) students and found that, after a 26-week treatment period (reading/reading-while-listening to 20 graded readers), learners in the reading-while-listening group outperformed those in the reading-only group. Similarly, Chang and Millet (2014) demonstrated the advantage of reading-while-listening over both reading-only and listening-only for the development of listening skills. The low-intermediate EFL learners in their study read and/or listened to 10 graded readers over a 13-week period and, while all conditions led to improvements in listening skills, the reading-while-listening group had the highest comprehension scores. Compared to reading-only and listening-only, L2 adult learners also seem to prefer reading-while-listening (e.g., Brown, Waring, & Donkaewbua, 2008).

Several explanations have been proposed to describe the potential benefits of the addition of auditory input for reading comprehension. First, the audio in reading-while-listening might force learners to read faster, improving their reading rate and fluency (Chang & Millet, 2015). Second, the audio in reading-while-listening might help learners develop correspondences between the phonological and orthographic forms of words. In a recent eye-tracking study, Conklin et al., (2020) showed that L1 and L2 readers in reading-while-listening conditions read ahead of the audio. As the authors explain, the written text provides listeners with a visual cue for the boundaries of the forthcoming word that they are about to hear, which may help speed word segmentation. Third, the presence of written and auditory input in reading-while-listening allows learners to approach the task according to their strengths and preferences (Tragant, Muñoz, & Spada, 2016). Furthermore, learners have often reported their preference for the reading-while-listening modality (e.g., Brown, Waring, & Donkaewbua, 2008; Chang & Millet, 2014; Tragant, Muñoz, & Spada, 2016), which might also support learning.

While empirical evidence generally suggests the benefits of reading-while-listening for the development of a range of skills, its advantage over reading-only has not always been demonstrated. Similar improvements from reading-while-listening and reading-only were reported in relation to vocabulary learning (e.g., Brown, Waring & Donkaewbua, 2008), reading rate (Taguchi, Takayasu-Maass, & Gorsuch, 2004) and comprehension (e.g., Serrano & Pellicer-Sánchez, 2019; Pellicer-Sánchez, et al., 2020). Diao and Sweller's (2007) study also showed that EFL learners' comprehension from reading-while-listening was in fact lower than that of reading-only. In addition, the majority of these investigations have focused on the comparison of reading-while-listening and reading-only/listening-only conditions where only verbal input was presented (either written, aurally or both) (e.g., Brown, Waring, & Donkaewbua, 2008; Webb, 2012, 2015). The few studies that have examined the affordances of the addition of auditory input in the presence of non-verbal input (i.e., pictures) are reviewed in the following section.

Eye movements in multimodal reading

Jewitt, Bezemer and O'Halloran (2016) identify multimodal reception analysis as one of the key frameworks for studying multimodal texts. This approach uses eye-tracking to examine the cognitive processes that underlie the reception and processing of multimodal texts. Thus, the use of eye-tracking in multimodal and multimedia learning allows researchers to obtain a direct indication of cognitive processing during multimedia learning (Mayer, 2017). As Holsanova (2014) explains, “eye movement data serve as a window on the mind of the user, revealing perceptual and cognitive processes underlying users’ interactions with the multimodal messages” (p. 293).

There is a growing body of empirical studies on the perception and integration of text and images in education research (Holsanova, 2014). As indicated previously, the vast majority of eye-tracking studies on multimedia and multimodal learning have been conducted in the L1, focusing on learning in a particular domain. In their meta-analysis of 58 eye-tracking studies on multimedia learning, Alemdag and Cagiltay (2018) found that most of the studies had been conducted on science materials. For example, Hannus and Hyönä (1999) examined 10-year-old elementary school children’s eye movements during learning from illustrated science textbook passages. They investigated how children divided their attention between text and illustrations. Results showed that learning was heavily driven by the text and that children inspected illustrations only minimally. Most eye-tracking studies examining attentional focus on text and pictures have supported this pattern, showing that learners tend to give more attention to text than pictures (e.g., Johnson & Mayer, 2012; Schmidt-Weigand, 2011; Schmidt-Weigand, Kohnert, & Glowalla, 2010). Schmidt-Weigand, Kohnert, and Glowalla (2010) found that, when presenting images with spoken text, learners spent more time on the images than when they were presented with images and written text. Previous studies have also examined the allocation of attention to static vs. dynamic visuals in multimedia environments and have shown that learners allocate more visual attention to dynamic images (e.g., Chen et al., 2015; Takacs & Bus, 2016). Notably, the integration of text and pictures has been shown to support retention and enhanced performance (e.g., Mason, Tornatora, & Pluchino, 2015).

Very few eye-tracking studies on multimodal reading have been conducted in the L2 learning context. Warren, Boers, Grimshaw, and Siyanova-Chanturia (2018) examined the effect of multimodal glosses in reading (compared to picture only and text only glosses) on learning and comprehension. They found that L2 adult learners paid less attention to the text when pictures were simultaneously presented in the glosses. Despite the processing differences, no effect was found on comprehension. Previous studies have also shown that, in line with results in the L1 context, young L2 learners also allocate more attention to multimedia materials that contain dynamic images than to materials containing static images (e.g., Tragant & Pellicer-Sánchez, 2019).

Two recent studies examined the effect of auditory input on young L2 learner’s engagement with text and pictures in multimedia materials (i.e., Pellicer-Sánchez, et al., 2020; Serrano & Pellicer-Sánchez, 2019). In these studies, young EFL learners were asked to read and to read-while-listen to an illustrated graded reader for comprehension. Analyses of eye movements showed that overall, learners spent more time on the text than on the pictures in both conditions and that more time was spent on the images in reading-while-listening conditions. Interestingly, these processing differences were not reflected in differences in comprehension. This research also indicated that more time reading the text was related to comprehension difficulties (Pellicer-Sánchez et al., 2020; Serrano & Pellicer-Sánchez, 2019), while more time looking at the images was associated with better comprehension (Pellicer-Sánchez, et al., 2020).

In these two studies, the relationship between processing patterns and comprehension may be the result of the young learners' developing literacy skills. Whether these patterns are representative of L2 multimodal reading more generally, or are characteristic of young learners' processing, needs to be determined. As Jewitt, Bezemer and O'Halloran (2016) argue, "By accumulating eye-tracking data from a large number of respondents in experimental settings, general patterns in the reception of multimodal artefacts can be explored" (129). However, the scarce eye-tracking evidence is restricted to studies with young EFL learners. Thus, our understanding of how L2 readers engage with different types of multimodal texts is still rather limited.

THE STUDY

Eye-tracking studies on multimodal and multimedia reading have shown that, when simultaneously presented with text and images, readers spend more time on the text than the images. Crucially, learners spend more time on images in reading-while-listening than in reading-only. Further, research has shown that the integration of text and pictures is crucial for multimodal comprehension, but little is known about how auditory input affects that integration. Previous studies have not examined the effect of auditory input on the processing and comprehension of multimodal materials by L2 adult readers, nor their relationship with comprehension. Research has not examined how (and if) L2 processing of multimodal materials differs from those of L1 readers. The present study aimed at addressing these gaps. The following questions were addressed:

1. Are there differences in adult readers' allocation of attention and integration of text and images in reading-only and reading-while-listening conditions?
2. Are patterns of attention allocation and integration of text and images different for L1 and L2 readers?

3. Is the processing of the text and pictures in multimodal materials related to comprehension?

To answer these questions, adult L1 and L2 readers were asked to read and to read-and-listen to an illustrated story while their eye movements were recorded. Comprehension of information present in both the text and the images was examined immediately after the reading. The relationship between processing patterns and performance measures was explored. Based on previous research findings, it was hypothesised that participants would allocate more attention to the text than images (e.g., Johnson & Mayer, 2012; Schmidt-Weigand, 2011) and that more time would be allocated to images in the presence of auditory input (e.g., Serrano & Pellicer-Sánchez, 2019; Pellicer-Sánchez et al., 2020). It was also hypothesised that processing time on text would be negatively related to comprehension (e.g., Serrano & Pellicer-Sánchez, 2019; Pellicer-Sánchez et al., 2020), whereas processing time on images would be positively related to comprehension (Pellicer-Sánchez et al., 2020).

METHODOLOGY

Participants

Forty-nine participants initially took part in the study. Data from two participants were discarded due to poor calibration and track loss. This resulted in a total of 47 participants, including 25 L2 (17 female; 8 male) and 22 L1 (19 female; 3 male) speakers of English. The L2 speakers were advanced learners of English studying at a higher education institution. They came from different L1 backgrounds and had met the entry requirement to study at the university. Their ages ranged from 19 to 29 (M = 22.24; SD = 2.74). Although collecting vocabulary size data from this group was not possible, it was important to have a good estimate of their vocabulary size so as to design materials that were appropriate for their proficiency. The vocabulary size of L2 speakers was estimated based on the performance of another group of participants with similar characteristics on the X_Lex vocabulary size test (Meara & Milton, 2003). The vocabulary size was estimated at a mean of 7777 words (Min = 6000, Max = 9666, SD = 1000). L1 speakers were all undergraduate students at the same institution majoring in a range of subjects. Their ages ranged from 19 to 25 (*M* = 19.45; *SD* = 1.57). Participation was voluntary and they all received a small compensation for their time.

Reading materials

A short narrative (2070 words) was created for the purposes of the study. A mystery story was written around a set of 31 images taken from the picture stimuli developed by the Arnold Lab (Rosa & Arnold, 2017; http://jaapstimuli.web.unc.edu/). The original stimuli consisted of 53 pairs of pictures which together depict a mystery story. Since the text was built around the images, the images had a close relationship with the text, supporting and representing the information that was presented in the text. More precisely, the images illustrated the characters and actions described in the story. These images were chosen because, while they were originally created with a different research purpose, they had been specifically designed for experimental purposes controlling for their visual properties. They were designed by the same illustrator and the same illustration technique was used throughout the pictures. The main visual features (i.e., shapes and colours used) remained constant across the images. The story was displayed over 31 pages. Each page contained a block of text and a picture. We controlled for the vocabulary used in the story to ensure that the text would not pose comprehension difficulties for the participants. The majority of the words in the text (96%) were from the first 6K most frequent words in English. Based on the proficiency of participants and their estimated vocabulary size of the participants, it was likely that they would be familiar with the first 6k most frequent words in English and that the content would not be too difficult for them. Two words which were beyond the 6K level (i.e., butler, chauffeur) were introduced to the participants at the beginning of the experiment, as they were crucial to understanding the narrative. The story was presented on a computer across 31 screens with similar amount of text: an average of 66.77 words per screen (SD = 8.18; Min = 51; Max = 84) (see Appendix A for a sample page). The text was presented in Times New Roman, point-25 font, with 1.5 line spacing in order to fill the text region of interest as much as possible, while keeping the size of the text region comparable to that of the image area.

Half of the story was presented in the reading-only mode and the other half in the reading-while-listening mode (i.e., a visual and auditory version of the story) in a within-subjects, counterbalanced design. The 31 screens were presented in a blocked design, with a first block of 16 trials and a second block of 15 trials. The position of the images relative to the text was counterbalanced to avoid a picture always being presented on the same side (i.e., to left or right of the screen). To account for the odd number of pages, two lists were created ("16left+15right", "15left+16right") and counterbalanced across the sample. Additionally, the relative positions were randomised for each individual trial.

The auditory input for the reading-while-listening was recorded by a male L1 speaker of English. The recording had an average speed of 3.12 words per second (with a typical speech rate being 3.7 words second for native speakers of English; Goldman-Eisler, 1961).

Comprehension test

Our aim was to gain a comprehensive understanding of multimedia reading and the role of both textual and pictorial input on multimodal comprehension. Thus, following Pellicer-Sánchez et al., (2020) two types of questions were designed, i.e., text-related questions and image-related questions. Text-related questions were those whose responses were included in the text and also supported by the images, whereas image-related questions were those related to specific details presented only in the images. For the text-related questions, each narrative section was first parsed into idea units, defined as distinct events or actions that occurred in the course of the story. Idea units were then explored as potential true/false and multiple-choice comprehension items. Two items were designed for each narrative section – one true/false and one multiple-choice item. The result of this process was 62 text-based comprehension items, which were piloted.

The pictures linked to the narrative were then examined to create image-based comprehension items. The first step was to identify elements in the pictures that were related to the narrative but not mentioned in the text. This was done to provide comprehension items that were linked solely to viewing the associated imagery. Eleven multiple-choice items from separate pictures were created based around these elements. This is less than the total number of pictures presented in the narrative because not every picture contained imagery that was explicitly linked to the overall narrative but not mentioned in the written text. The text-based and image-based items were combined into a 73-item comprehension test. The items were presented in the order they appeared in the narrative.

Results of the comprehension test were analysed to identify poorly performing items that needed to be removed from the final results. The reliability of the test items was examined by calculating Cronbach’s alpha for all test items. The comprehension items had an α of .816, which is considered to be acceptable reliability for a test of this nature and thus no items were removed or modified based on this measure. The reliability of the text-related questions was acceptable (α = .816) but, as expected, the reliability of the image-related questions was not acceptable (α = .329) due to the small number of items included (N = 11). Thus, item discrimination was carried out as a second analysis on the comprehension test. The discrimination index for text-based items and image-based items were calculated separately. For the image-based items, the participants were split into a high score and a low score group. For the text-based items, the participants were split into the top one-third and the bottom one-third as the scores on these items lent themselves better to this approach to participant separation. Both approaches are considered acceptable in the test design literature. According to this analysis, none of the image-based items were cause for concern. For the text-based items, based on their negative item discrimination scores, two items were removed. The final comprehension results for this study were then based on the remaining 71 items (60 text-based and 11 image-based). This revised set of items had an α of .827 which was a slight improvement on the original analysis (see Appendix B for sample items).

Procedure

Data were collected individually in a psycholinguistics laboratory. Participants first read the information sheet and provided their written consent. L2 speakers also completed a language background questionnaire. After setting up the equipment, the experimenter went through the instructions with the participant. They were asked to read and read-while-listening to a short story as naturally as possible for comprehension. Participants were aware of the forthcoming comprehension test. Headphones were provided and were worn throughout the experiment as they aided concentration and blocked out noise.

The narrative was presented on a widescreen monitor at a resolution of 1920x1080 to better accommodate text and images arranged horizontally. In the reading-only condition, pages advanced with a mouse click, whereas in the reading-while-listening condition the pages advanced automatically when the audio recording finished. This different pacing of reading-only and reading-while listening more closely resembles classroom practices. Eye tracking was performed using an SR Research EyeLink 1000+ at a sampling rate of 1000 Hz, with an accuracy of 0.25-0.5°, and a precision of < 0.01°. Recording was monocular (right eye), and in the head-stabilised mode. A 9-point calibration was conducted at the beginning and half-way through the experiment (before the change of modality). A drift correction was also performed after each screen and additional calibrations were carried out when the experimenter considered them necessary. After the eye-tracking experiment, participants were asked to complete the comprehension test. The whole procedure lasted on average 45 minutes.

Analysis

The comprehension test was scored following a dichotomous system (1 for correct responses and 0 for incorrect responses). For the analysis of eye movements, two regions of interest of equal area (488235 px^2) were defined for each trial, surrounding the image and the block of text. Two types of eye movements were explored, i.e., fixations (i.e., when visual gaze is maintained on a single location) and saccades (i.e., movements of the eyes). Fixations shorter than 80 ms were removed from the dataset (4.62% of fixations), but other fixations were not merged. The following eye-movement measures were extracted and analysed:

- Dwell time % (the percentage of the sum of all fixation durations within each region of interest)
- Fixation % (the percentage of the total number of fixations in a trial within each region of interest)
- Average fixation duration within each region of interest
- Number of integrative saccades between text and images

Dwell time %, and fixation % were chosen as measures of allocation of attention. Average fixation duration was selected as it has been claimed to be a good measure of how attention unfolds over time. Finally, it has been claimed that frequent transitions between text and images are indicative of integration (Arndt, Schüler, & Scheiter, 2015; Hegarty & Just, 1993; Holsanova, 2014; Scheiter & Eitel, 2017). Thus, the number of integrative saccades between text and images was chosen as a measure of integration.

Data were analysed via linear mixed-effect models using the *lme4* (v1.1-21; Bates, Mächler, Bolker & Walker, 2015) package for *R* (v 3.6.1 R Core Team, 2019). Post-hoc comparisons were carried out using the *multcomp* package (v1.4-13; Hothorn, Bretz & Westfall, 2008). Models included random intercepts for participants (detailed information about the models is provided in the next section).

RESULTS

In response to the first and second research questions, participants' eye movements to the two regions of interest (i.e., text and image areas) in the two modalities (i.e., reading-only and reading-while listening) were explored. Dwell time descriptive statistics decomposed by the different factors considered in the analysis can be found in Table 1. Because the duration of the trials (and hence the total dwell time) was limited by the duration of the audio recordings in the reading-while-listening condition whereas reading in reading-only trials was self-paced, percentage measures were entered in the models as a way of controlling for differences in trial length.

TABLE 1

Dwell Time Descriptive Statistics by Condition (Reading-only, Reading-while-listening) and Type of Region (picture, text). SD in Brackets

Group	Condition	Region	Dwell Time	Dwell Time %
L1 speakers	Reading-only	Picture	1285 ms (85)	9.21% (0.4)
	Reading-only	Text	11386 ms (215)	89.11% (0.4)
	Reading-while-listening	Picture	1764 ms (83)	11.05% (0.5)
	Reading-while-Listening	Text	14629 ms (165)	87.86% (0.5)
L2 speakers	Reading-only	Picture	1412 ms (65)	7.89% (0.3)
	Reading-only	Text	16128 ms (283)	90.53% (0.3)

Reading-while-listening	Picture	1789 ms (95)	11.00% (0.5)
Reading-while-listening	Text	14262 ms (145)	87.41% (0.5)

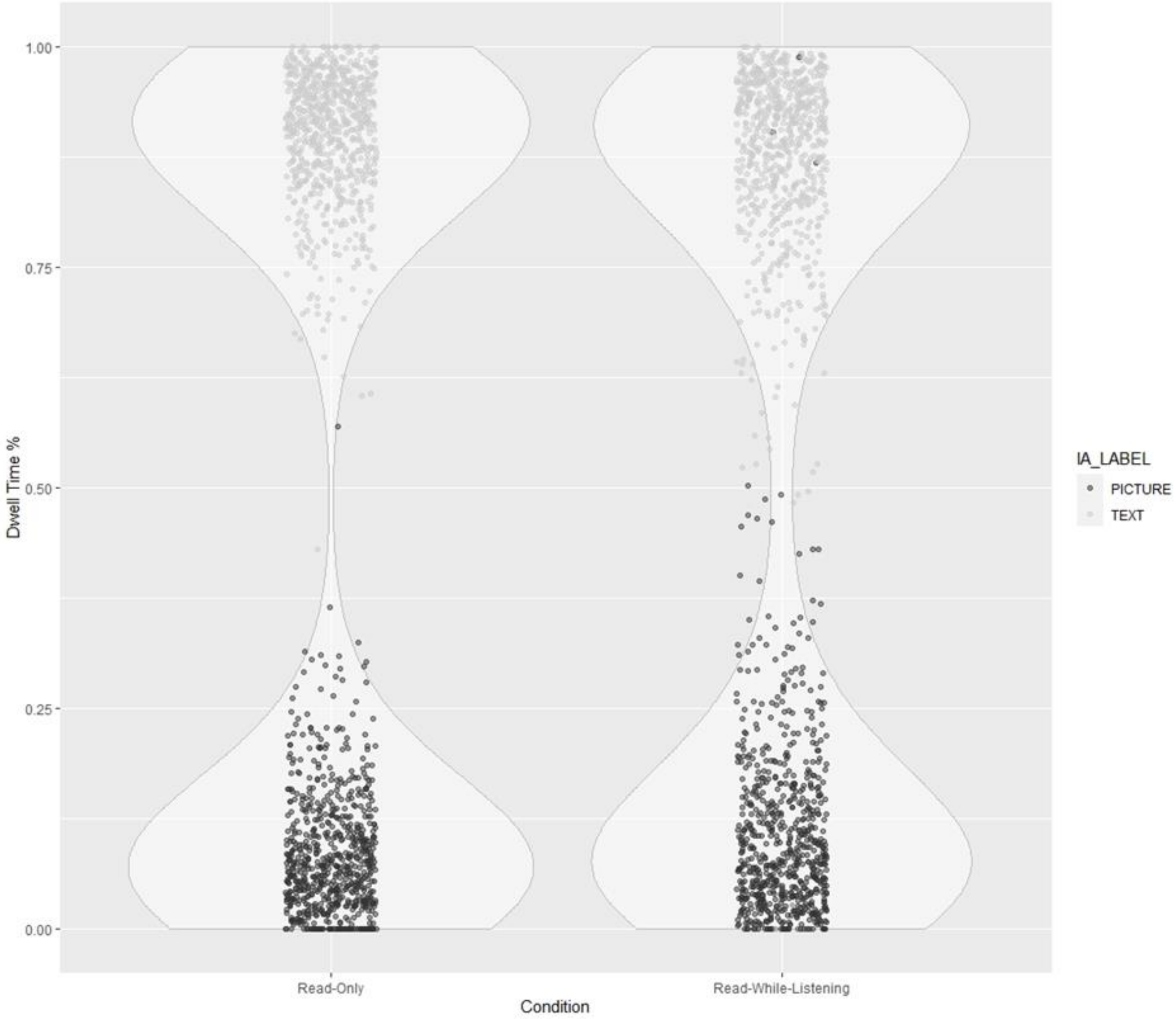
Two model structures were fitted to the data: the first with main effects of and interactions between Condition and Region, and the second with the addition of Group as a factor. Both structures included random intercepts for participants.

m1) IA_DT_% ~ CONDITION * REGION + (1 | PARTICIPANT)
m2) IA_DT_% ~ GROUP * CONDITION * REGION + (1 | PARTICIPANT)

The two models were compared via maximum-likelihood (ML). The addition of Group as a factor did not increase the goodness of fit of Model 2 over Model 1 (see Appendix C). The latter revealed significant main effects of both condition, $\beta = 0.02$, $t(2910) = 5.31$, $p < .0001$, $\eta^2_p = .009$, $d = 0.20$, and region, $\beta = 0.81$, $t(2910) = 169.76$, $p < .0001$, $\eta^2_p = .91$, $d = 6.29$, as well as a significant interaction between the two factors, $\beta = -0.04$, $t(2910) = -7.13$, $p < .0001$, $\eta^2_p = .02$, $d = 0.26$. To decompose the interaction (Figure 1), Tukey contrasts for multiple comparisons with Bonferroni-Holm correction were performed between all levels of the two factors. These revealed that the text region was fixated significantly more during reading-only trials than during reading-while-listening trials, $\beta = -0.02$, $z = -4.77$, $p < .0001$, $d = 1.3$, whereas the images were fixated significantly longer during reading-while-listening trials, $\beta = 0.02$, $z = 5.31$, $p < .0001$, $d = 1.55$.

FIGURE 1

Interaction Between Condition (Reading-only vs. Reading-while-listening) and Region (Text vs. Pictures) for the Dwell Time % Measure



A similar analysis was performed on the number of fixations recorded within the regions of interest, and, as with Dwell Time, these were computed as percentages of the total number of fixations recorded during each trial. Table 2 reports the descriptive statistics for both counts and percentages. The same model structures were fitted to this dependent variable, and Model 1 was again found to be a better fit to the data (see Appendix C). This model revealed significant main effects of both condition, β = 0.02, *t*(2910) = 6.01, *p* < .0001, η^2_p = .01, *d* = 0.22, and region, β = 0.82, *t*(2910) = 196.64, *p* < .0001, η^2_p = .93, *d* = 7.29, as well as a significant interaction between the two factors, β = -0.04, $t(2910) = -8.37$, $p < .0001$, $\eta^2_p = .02$, $d = 0.31$. The interaction suggests that participants made more fixations on the images during reading-while-listening trials than in reading-only trials, β = 0.02, $z = 6.01$, $p < .0001$, $d = 1.75$. Similarly, they made significantly more fixations on the text during reading-only than in reading-while-listening trials, β = -0.02, $z = -5.83$, $p < .0001$, $d = 1.70$ (Figure 2).

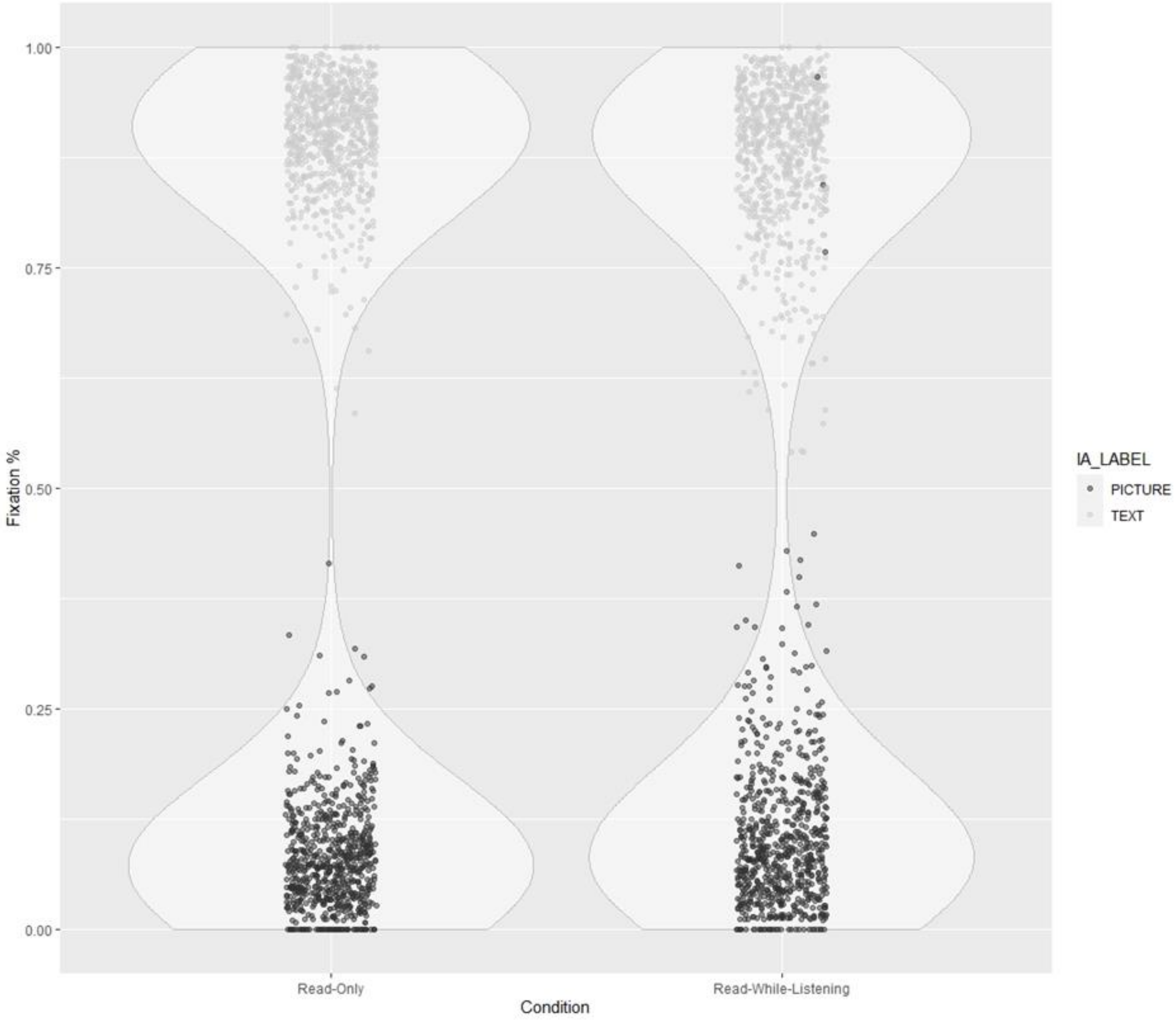
TABLE 2

Fixation Count Descriptive Statistics by Condition (Reading-only, Reading-while-listening) and Type of Region (picture, text). SD in Brackets

Group	Condition	Region	Fixations N	Fixations %
L1 speakers	Reading-only	Picture	5.73 (0.30)	8.57% (0.3)
	Reading-only	Text	56.16 (0.94)	89.45% (0.3)
	Reading-while-listening	Picture	7.82 (0.34)	10.53% (0.4)
	Reading-while-listening	Text	64.33 (0.64)	87.93% (0.4)
L2 speakers	Reading-only	Picture	6.17 (0.24)	7.53% (0.2)
	Reading-only	Text	75.15 (1.23)	90.59% (0.2)
	Reading-while-listening	Picture	7.51 (0.31)	10.48% (0.4)
	Reading-while-listening	Text	64.04 (0.63)	87.42% (0.5)

FIGURE 2

Interaction Between Condition (Reading-only vs. Reading-while-listening) and Region (Text vs. Pictures) for the Fixation Count % Measure



We then proceeded to analyse the average duration of fixations in the two regions of interest. Table 3 includes the descriptive statistics for this measure. In this case, the model structure that included Group as a factor was found to better fit the data than the simpler structure (See Appendix C). Despite this, the model did not reveal significant effects of participant group, but rather a significant main effect of condition, β = 12.97, *t*(2767) = 3.38, *p* = .0007, η^2_p = .004, *d* = 0.13, and a significant interaction between condition and region, β = 13.55, *t*(2764) =2.56, *p*

= .01, η^2_p = .0006, *d* = 0.10. The latter indicates that the difference between the two conditions was more marked for the text region than for the picture region. The difference in average fixation duration between reading-while-listening and reading-only trials was significant for the pictures, β = 8.95, *z* = 3.54, *p* = .001, *d* = 1.03, with longer fixations on average during reading-while-listening trials. However, this difference was more significant for the text regions, which contained longer mean fixations during reading-while-listening trials than during reading-only trials, β = 17.69, *z* = 7.25, *p* < .0001, *d* = 2.12 (see Figure 3 for a plot of this effect).

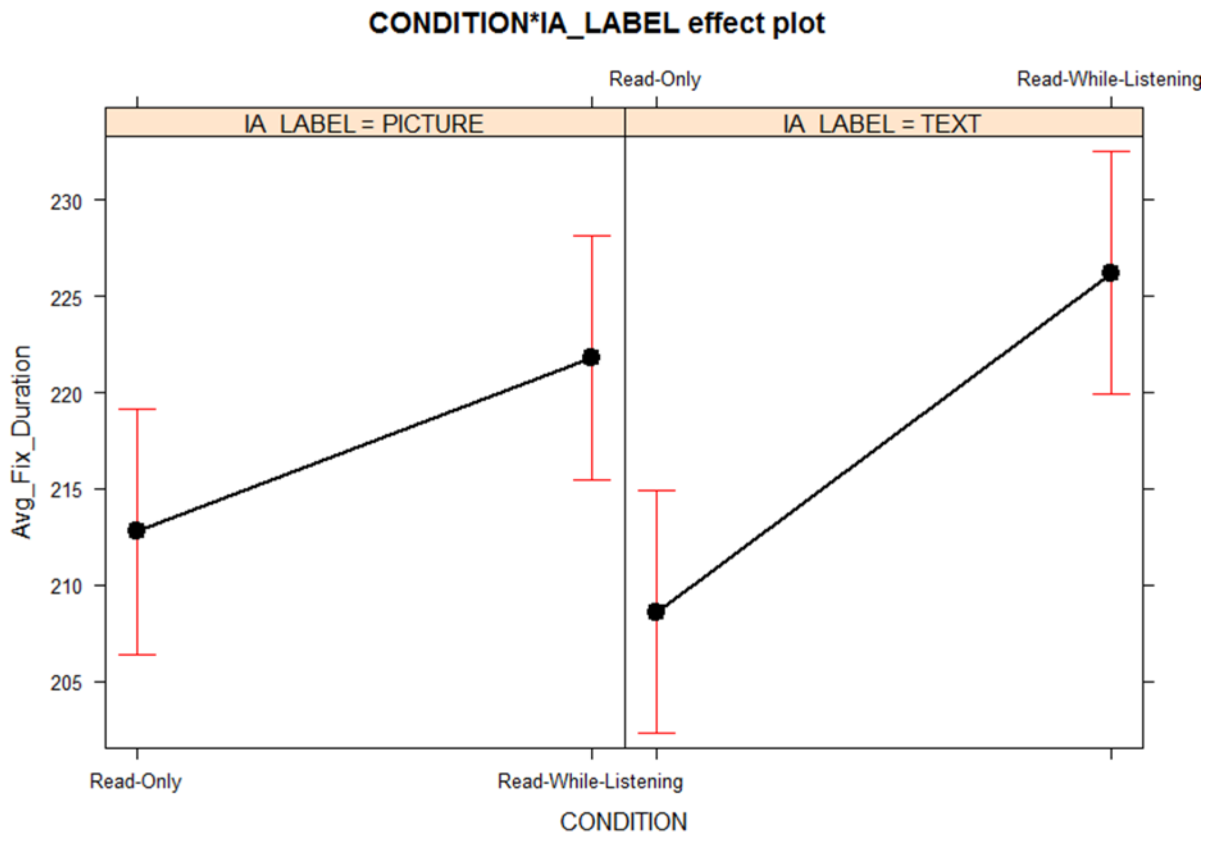
TABLE 3

Average Fixation Duration Descriptive Statistics by Condition (Reading-only, Reading-while-listening) and Type of Region (picture, text). SD in Brackets

Group	Condition	Region	Fixation Duration
L1 readers	Reading-only	Picture	209 ms (3.32)
	Reading-only	Text	202 ms (1.29)
	Reading-while-listening	Picture	220 ms (3.90)
	Reading-while-listening	Text	228 ms (1.92)
L2 readers	Reading-only	Picture	217 ms (3.51)
	Reading-only	Text	213 ms (1.12)
	Reading-while-listening	Picture	222 ms (3.47)
	Reading-while-listening	Text	223 ms (1.33)

FIGURE 3

Plot for the Interaction Between Condition (Reading-only vs. Reading-while-listening) and Region (Text and Picture) for Average Fixation Duration



Subsequently, we analysed the number of integrative saccades out of all saccades recorded in each trial. These are reported in Table 4. For this purpose, we fit a generalised linear mixed-effect model (using *lme4::glmer*) to predict saccade outcomes (i.e., whether they were integrative or not) based on experimental condition and participant group. The model used a binomial family function with a logit link function.

The following model structures were fitted to the data:

m1) INTEGR_YES_NO ~ CONDITION + (1 | PARTICIPANT)

m2) INTEGR_YES_NO ~ GROUP * CONDITION + (1 | PARTICIPANT)

Model comparison revealed that the addition of the Group factor did not increase the goodness of fit of the model (see Appendix C), so Model 1 was used. The main effect of condition was found to be significant, β = 0.41, *z* = 14.35, *p* < .0001, *d* = 4.19, indicating a significantly higher proportion of integrative saccades during reading-while-listening trials compared to reading-only trials.

TABLE 4

Descriptive Statistics for Integrative Saccades by Condition (Reading-only vs. Reading-while-listening) and Type of Region (Picture vs. Text). SD in Brackets

Group	Condition	Integrative Saccades %	Integrative Saccades N
L1 speakers	Reading-only	4.21% (0.16)	883
	Reading-while-listening	5.83% (0.22)	1417
L2 speakers	Reading-only	4.00% (0.13)	1297
	Reading-while-listening	6.06% (0.22)	1805

In response to the third research question, we attempted to determine whether eye movement measures could be used as predictors of participants' behavioural performance in the comprehension post-test, which contained questions specific to either the content of the text or the images. We first computed distinct accuracy percentages for text-related and image-related questions, and we further decomposed these into accuracy percentages on question probing information presented during reading-only and reading-while-listening trials for each participant. Mean accuracy was computed on this basis, and the same pre-processing stages were performed on the Dwell Time percentage measure. Table 5 reports the descriptive statistics for response accuracy.

TABLE 5

Descriptive Statistics for Response Accuracy on the Comprehension Test by Condition (Reading-only, Reading-while-listening) and Type of Question (Image-related, Text-related)

Group	Condition	Image % Correct[a]	Text % Correct[a]
L1 readers	Reading-only	45.5% (4.9)	72.5% (3.2)
	Reading-while-listening	47.3% (4.5)	69.5% (2.7)
L2 readers	Reading-only	57.4% (3.4)	80.6% (2.1)
	Reading-while-listening	49.8% (3.6)	76.2% (2.3)

[a] Guessing probability was 33% for the multiple-choice items (considering only the possible options) and 25% when considering all options (including “I don’t know”), and 50% for the True-False items. Combined guessing probability (multiple-choice and True-False items) from the possible options was 41.5%, and 37.5% from all the options presented.

Before attempting to relate eye-tracking measures and behavioural measures we fitted the following two models to the response accuracy data to determine whether accuracy in responding to questions pertaining to the text or to the images was significantly different between participant groups and conditions.

m0) Image_Accuracy ~ CONDITION * GROUP + (1 | PARTICIPANT)

m0_2) Text_Accuracy ~ CONDITION * GROUP + (1 | PARTICIPANT)

In both cases, only a significant main effect of participant group was observed (β = 0.11, $t(86.44) = 2.03$, $p = .04$, $\eta^2_p = .05$, $d = 0.44$, and $\beta = 0.08$, $t(66.99) = 2.17$, $p = .03$, $\eta^2_p = .07$, $d = 0.53$ respectively), suggesting generally better performance on the part of L2 readers.

The following linear mixed models were subsequently fitted to the data, with the goal of determining whether response accuracy could be predicted by the degree of attention allocated to the relevant elements of the stimuli:

m1) Image_Accuracy ~ avg_DT_%_IMAGE * CONDITION * GROUP + (1 | PARTICIPANT)

m2) Text_Accuracy ~ avg_DT_%_TEXT * CONDITION * GROUP + (1 | PARTICIPANT)

Model 1 revealed significant main effects of the Dwell Time percentage on the images, β = 2.02, *t*(85.76) = 2.59, *p* = .01, η^2_p = .07, *d* = 0.56, and of participant group, β = 0.28, *t(*85.98) = 2.48, *p* = .01, η^2_p = .07, *d* = 0.53, but not of condition. This indicates a general trend whereby more time spent looking at the images predicts greater accuracy when answering questions concerning their content. The effect of participant group, on the other hand, indicates that the L2 group achieved greater mean accuracy in answering these questions. However, this was likely due to the floor performance of two L1 readers, and further investigation would be required to ascertain the practical importance of this effect.

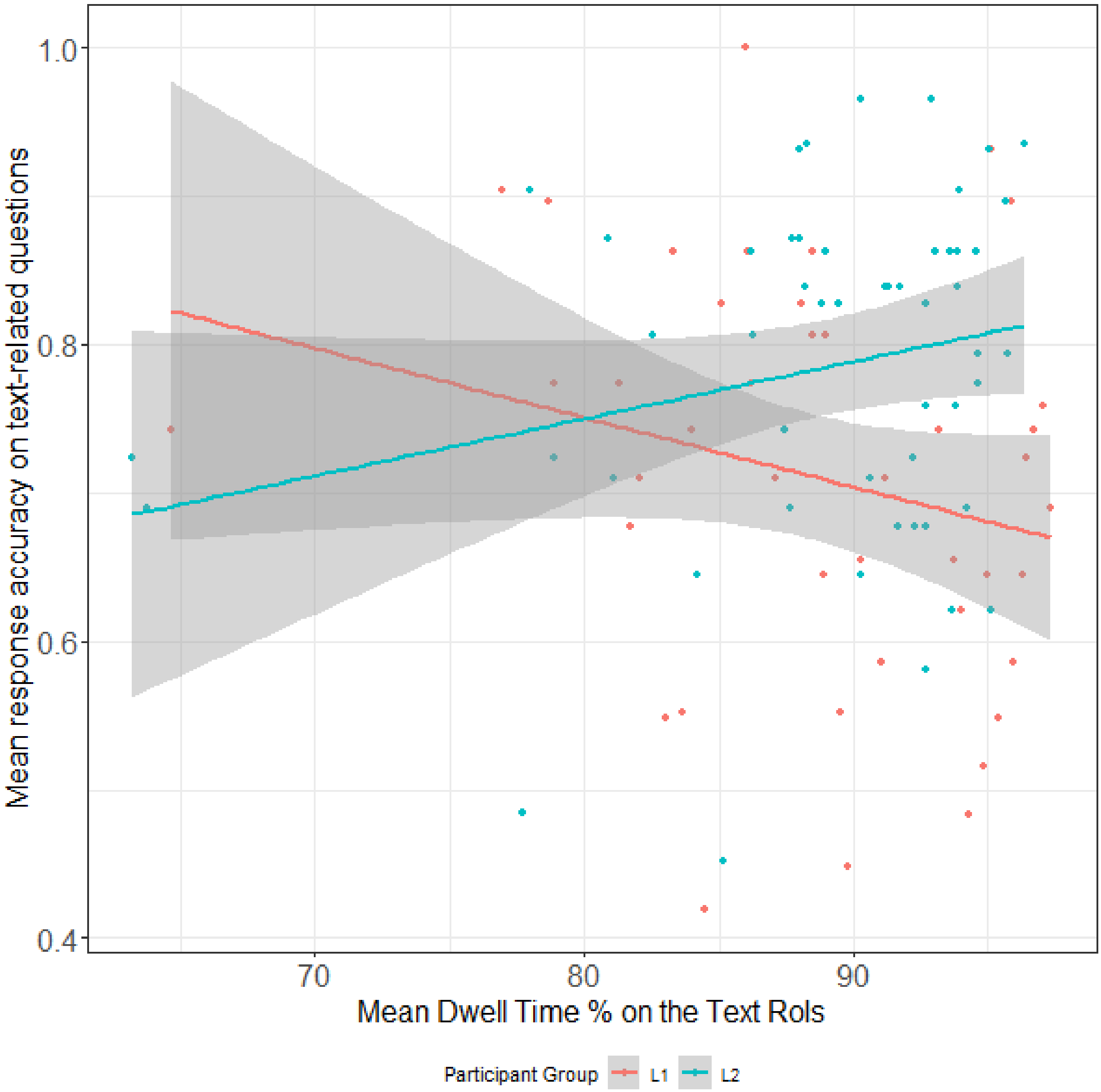
Fitting the same model to the text-related questions (m2) yielded a different pattern of results. Namely, a significant main effect of participant group, β = -1.23, *t*(73.86) = -2.33, *p* = .02, η^2_p = .07, *d* = 0.54, and a significant interaction between Dwell Time percentage on the text and participant group, β = 1.46, *t*(73.59) = 2.48, *p* = .01, η^2_p = .08, *d* = 0.58. A post-hoc analysis of this interaction produced only a significant difference between L1 and L2, β = -0.06, *t*(44.5) = -1.95, *p* = .05. Nevertheless, an inspection of the interaction plot (Figure 4) indicates a distinct relationship between average Dwell Time percentage and response accuracy for the text-related questions between L1 and L2 speakers. Namely, this interaction would seem to suggest that fixating the text for longer is related to increased response accuracy for L2 readers but is indicative of processing difficulties for L1 speakers.

The following final model structure was also fitted to the data:

m3) Text_Accuracy ~ avg_DT_%_IMAGE * CONDITION * GROUP + (1 | PARTICIPANT)

FIGURE 4

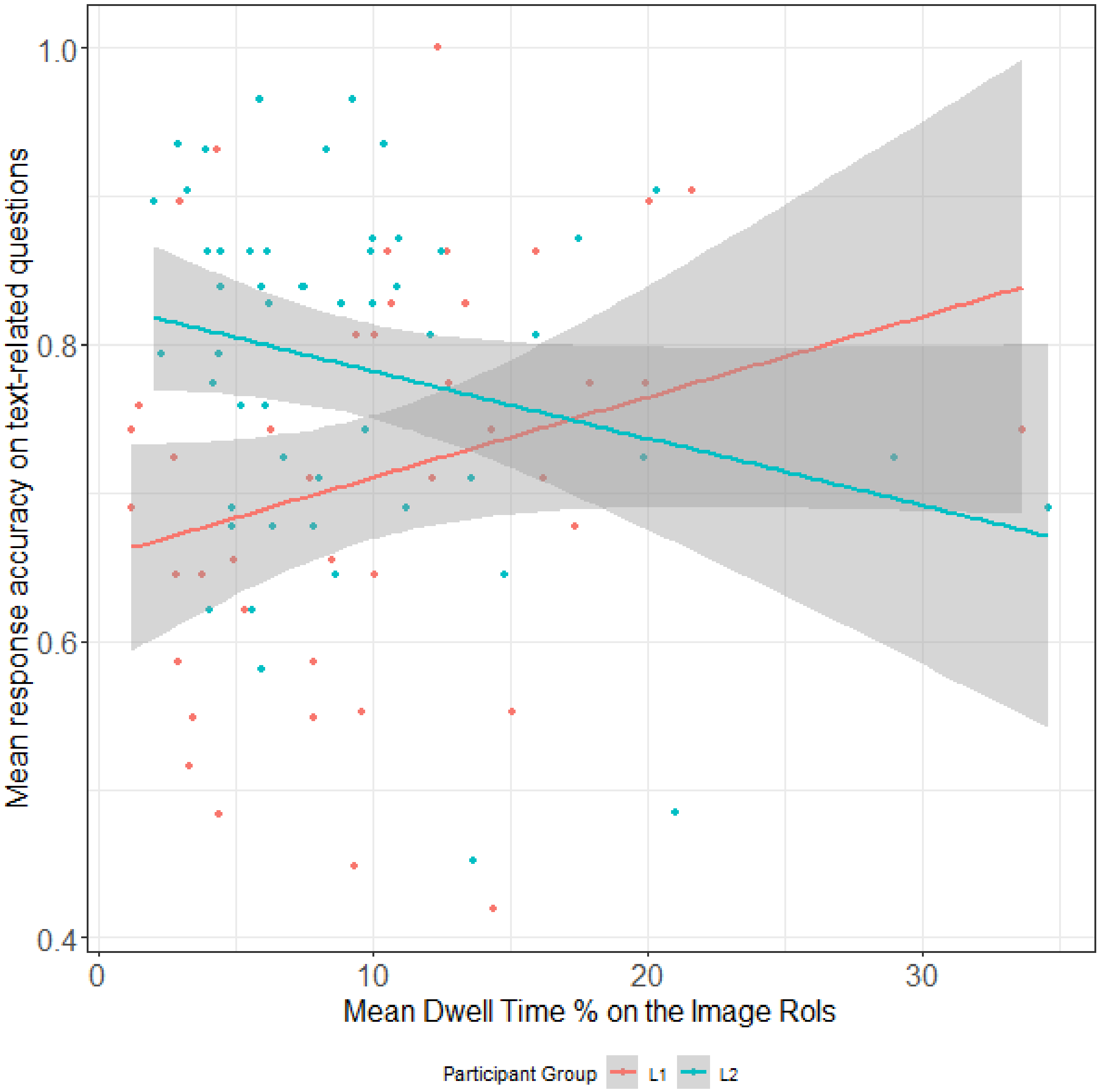
Plot of the relationship between Mean Dwell Time % on the text regions and participant's mean response accuracy on text-related questions for the two participant groups. Shaded areas represent 95% confidence intervals.



The aim of the final analysis was to determine whether time spent observing the image (and, therefore, processing the information contained in it) facilitated the processing of the text itself, thereby improving response accuracy for text-related questions. This model produced a significant main effect of participant group, β = 0.21, *t*(85.33) = 3.44, *p* < .001, η^2_p = .12, *d* = 0.74, a significant interaction between Dwell Time percentage on the images and participant group, β = -1.70, *t*(73.58) = -2.73, *p* = .007, η^2_p = .09, *d* = 0.64, and a significant three-way interaction between Dwell Time percentage on the images, condition, and participant group, β = 1.26, *t*(51.33) = 2.03, *p* = .04, η^2_p = .07, *d* = 0.57. Post-hoc comparisons revealed non-significant differences (all *ps* > .05), possibly due to insufficient group sizes. Visual inspection of the interaction plot (Figure 5) shows distinct relationships between Dwell Time percentage on the images and response accuracy for text-related questions between L1 and L2 speakers. This pattern of results suggests that, while processing non-verbal information can support the processing of related textual information, this is true only for L1 readers; the textual comprehension of L2 readers does not appear to be improved by longer processing of the accompanying images in either reading-only or reading-while-listening trials but is in fact worsened as a result.

FIGURE 5

Plot of the relationship between Mean Dwell Time % on the images and participant's mean response accuracy on text-related questions for the two participant groups. Shaded areas represent 95% confidence intervals.



DISCUSSION

Current L2 reading practices are largely multimodal, often involving the presentation of printed text, images and auditory input. The main aim of the present study was to examine the effect that the addition of auditory input had on adult L2 readers' allocation of attention to the text and pictures in multimodal reading, as well as the relationship between processing patterns and comprehension. L2 processing patterns were also compared to those of L1 adult readers. To this end we examined the amount of attention (Dwell time %, Fixation count %, and Average fixation duration) that L1 and L2 adult readers allocated to the text and pictures in an illustrated story in two conditions, i.e., reading-only and reading-while-listening. In order to examine the integration of text and pictures in these conditions, we also examined the proportion of integrative saccades made between the text and image areas.

Results of the analysis of eye movements showed that in general L2 learners spent proportionally more time on the text than on the pictures in both conditions. This supports earlier findings with both L1 readers in domain learning (e.g., Alemdag & Cagiltay, 2018); Hannus & Hyönä, 1999; Johnson & Mayer, 2012; Schmidt-Weigand, 2011; Schmidt-Weigand, Kohnert, & Glowalla, 2010) and with young L2 language learners (e.g., Pellicer-Sánchez, et al., 2020; Serrano & Pellicer-Sánchez, 2019; Tragant & Pellicer-Sánchez, 2019). When engaging with multimodal materials, both L1 and L2 readers of different proficiencies tend to give more attention to the text, at least with the type of texts explored in these studies.

In response to the first research question, the presence of auditory input in reading-while-listening led to significant changes in processing patterns. Learners spent proportionally more time on the images in the reading-while-listening than in reading-only. Conversely, more attention was allocated to the text in reading-only conditions. The same pattern was found for the two measures of attention allocation (i.e., dwell time % and fixation count %). This finding is consistent with results of earlier studies with young L2 learners (e.g., Pellicer-Sánchez, et al., 2020; Serrano & Pellicer-Sánchez, 2019), and seems to suggest that, despite differences in age and language proficiency, the allocation of attention in this type of multimodal materials follows the same pattern. It is important to note that the effect sizes of the interactions between dwell time % and fixation count % were small (*d* = 0.2-0.3), possibly due to somewhat small group sizes. Interestingly, in response to the second research question, no main effect of group was found, suggesting that L2 adult readers processed text and pictures in a manner similar to L1 readers.

Importantly, the integration of text and pictures has been claimed to be key in multimodal reading comprehension. The present study indicates how this integration can be supported. A higher proportion of integrative saccades was made in the reading-while-listening condition, with a large effect of condition, demonstrating that the addition of auditory input supports the integration of text and pictures in multimodal materials. A similar tentative conclusion was made by Pellicer-Sánchez, et al. (2020). However, in their study this conclusion was based on measures of attention allocation, i.e., fixation count and dwell time percentages. According to the authors, the higher proportion of time spent processing the images in the reading-while-listening condition suggested a better integration of text and pictures. In the current study, we used measures that have been identified in the literature as specific measures of integration, i.e., integrative saccades. Thus, the current results provide stronger evidence for the value of reading-while-listening for the integration of text and images.

Regarding average fixation durations, results showed that overall fixations were longer in reading-while-listening than in reading-only, particularly in the text areas. This supports what has been suggested for L1 readers (Rayner, 1998). Interestingly, our average fixation durations for the text and images fall outside what has been suggested in the literature. Mean fixation durations in silent reading in the L1 range from 225-250 ms (Castelhano & Rayner, 2008). Longer average fixations in silent reading have been reported for L2 readers, with a mean fixation duration of 332 ms (Conklin, et al., 2020). However, our average fixation durations are shorter than these reported figures (L1 readers = 202 ms; L2 readers = 213 ms). Conklin et al. (2020) showed that in reading-while-listening, L1 and L2 readers had mean fixation durations of 323 ms and 325 ms respectively, falling within the range reported for oral reading, i.e., 275-325 ms (Castelhano & Rayner, 2008). Our average fixation durations are again shorter than these figures (L1 readers = 228 ms; L2 readers = 223 ms). Notably, while fixation durations on images are expected to be longer than on text (e.g., Rayner, 1998, 2009; Whitford & Joanisse, 2018), no main effect of region was found in the present study, suggesting very similar fixation durations on both texts and images. A potential explanation for these conflicting findings is that earlier investigations examined the processing of images/text on their own, as opposed to them being presented simultaneously, which was the case in the present study. It is important to consider that the results of average fixation durations should be interpreted with caution, as the effect sizes were very small.

Interestingly, despite the clear processing differences across conditions, there were no differences in comprehension. This means that the addition of auditory input did not seem to benefit comprehension. This supports findings of studies that failed to find a difference between reading-only and reading-while-listening conditions (e.g., Brown, Waring & Donkaewbua, 2008; Pellicer-Sánchez, et al., 2020; Taguchi, Takayasu-Maass, & Gorsuch, 2004; Serrano & Pellicer-Sánchez, 2019). These results have important implications for the applicability of the Cognitive Theory of Multimedia Learning to the L2 context. The present study did not show a detrimental effect of the simultaneous presentation of redundant written and aural input, going against the predictions of the redundancy principle and providing further evidence for its lack of applicability in the L2 learning context (e.g., Farias et al., 2014; Plass & Jones, 2005). In the reading-only condition, participants engaged with the pictures less. This may indicate less simultaneous activation of the verbal and non-verbal systems, i.e., less dual coding. In the reading-while-listening condition, while participants encountered redundant information via the audio and written text, they engaged more with the pictorial information. This should have allowed for a greater degree of dual coding. Thus, the inclusion of the pictorial information may have supported comprehension and overridden any potential detrimental effects of the redundant information. It should be noted that, in the absence of a more relevant theoretical account and following earlier studies on L2 learning from multimedia input, results are interpreted in light of the Cognitive Theory of Multimedia learning. However, as indicated earlier, the learning context examined in the studies that have confirmed the multimedia principles (science learning) is rather different from the type of learning and materials investigated in the present study (general comprehension). The relationship between pictures and text in the present study was different from that of studies on science learning, which is likely to explain the inapplicability of such principles. In addition, a comprehensive evaluation of the redundancy principle would involve a group that is exposed only to the auditory text and pictures.

A final aim of the present study was to explore the relationship between the processing of the text and image areas and comprehension, responding to the call for further research on the relationship between eye movements and performance measures (Alemdag & Cagiltay, 2018). To achieve a comprehensive understanding of the relationship between eye movement patterns and comprehension, we looked at both comprehension of information presented in the text (and supported by the images) and comprehension of details that were only presented in the images. This allowed us to further examine the role of pictures in multimodal reading. Concerning image-related questions, for both participant groups, longer time on the images was related to greater accuracy when responding to image-related questions. Interestingly, in Pellicer-Sánchez, et al. (2020), more attention to the pictures did not support the comprehension of image-related questions. The authors interpreted this finding as evidence that, when processing pictures, young L2 learners did not pay attention to the specific visual details that the questions addressed but that they used the images mainly to support comprehension. The comparison of their findings with those of the present study would suggest that adult readers process images differently, paying more attention to the visual characteristics of the images. Their more fully developed reading skills might allow them to allocate more cognitive resources to a deeper processing of images. It is worth noting that these results are based on responses to a small number of image-related questions (N = 11), with a very low reliability, and thus results should be treated with caution.

Regarding the time allocated to the text, results showed different effects for L1 and L2 speakers. For L1 speakers, a higher proportion of time on the text was negatively related to comprehension. This is in line with previous findings suggesting that longer time on the text might signal processing difficulties (e.g., Chang & Choi, 2014; Serrano & Pellicer-Sánchez, 2019). However, for L2 readers, a higher proportion of time on the text was positively related to comprehension of text-related questions. Conversely, a higher proportion of attention to the images supported comprehension of text-related questions for L1 speakers but not for L2 learners. These findings, together with results of previous studies, seem to suggest that increased attention to the text by both young L2 learners (Pellicer-Sánchez, et al., 2020; Serrano & Pellicer-Sánchez, 2019) and L1 readers seem to reflect processing difficulties, while it supports comprehension in the case of L2 readers. Further research with larger samples should be conducted to confirm these patterns.

Finally, results of the present study have also shown that, while there were no differences in the allocation of attention to text and pictures by L1 and L2 speakers, L2 readers achieved a higher level of comprehension. This is probably due to differences in how L2 readers approached the task and may suggest that the L2 readers were more accustomed to completing this type of language test. The better performance of L2 participants could also be attributed to age differences, as the mean age of the L2 group (*M* = 22.24, *SD* = 2.74; *p* < .0001) was significantly higher than the L1 group (*M* = 19.45, *SD* = 1.57, *p* < .0001).

It is important to note that the results of the present study are limited to advanced L2 readers. The similar patterns found for L1 and L2 readers are likely due to the advanced level of proficiency of L2 participants. Comparing our results to those of earlier studies with young learners provides an understanding of potential differences and similarities in processing patterns by proficiency level. However, more research needs to be carried out with L2 learners of different proficiencies to gain a comprehensive understanding of L2 learners' engagement with multimodal reading. It is also important to bear in mind that the processing patterns reported here are limited to multimodal texts where most of the information is presented in the text. Future studies need to explore other types of multimodal and multimedia materials. Finally, our examination of the role of pictures in comprehension is limited to a small number of image-related questions. To more fully understand the role that images play in text comprehension, research is needed on materials that allow for a larger number of image-related questions.

The results of the present study have important pedagogical implications. The findings confirm the value of adding auditory input for the integration of text and images in this type of multimodal narrative, without a detrimental effect on comprehension. Importantly, better integration has been shown to enhance retention and performance (e.g., Mason, Tornatora, & Pluchino, 2015). Previous research has also shown that adult readers seem to prefer reading-while-listening compared to reading-only and listening-only (e.g., Brown, Waring, & Donkaewbua, 2008). Taken together, these results support the use of reading-while-listening to illustrated texts in the L2 learning context. In addition, the results of this study have shown that adult readers pay attention to the details presented in images in illustrated texts, as confirmed by the relationship between processing time on images and accuracy on image-related questions. This finding further demonstrates that the potential of images for L2 learning should be exploited.

CONCLUSIONS

This study contributes to our understanding of L2 learners' integration of text and pictures in multimodal reading. Results show that, when engaging with multimodal texts, readers allocate most of their attention to the text. Crucially, the presence of auditory input leads to more integrative transitions between text and pictures in multimodal reading. Despite clear processing differences, similar levels of comprehension are obtained in the presence and absence of auditory input. L2 processing patterns are very similar to those of L1 readers. The study has also revealed that more attention to the text is negatively related to comprehension for L1 readers, while it seems to support comprehension for L2 learners, pointing to interesting differences between L1 and L2 readers in the relationship between eye movements and performance measures.

REFERENCES

Abraham, P., & Farías, M. (2017). Reading with eyes wide open: Reflections on the impact of multimodal texts on second language reading. *Íkala, Revista de Lenguaje y Cultura, 22*(1), 57–70.

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education, 125*, 413–428. doi: 10.1016/j.compedu.2018.06.023

Arndt, J., Schüler, A., & Scheiter, K. (2015). Text–Picture integration: How delayed testing moderates recognition of pictorial information in multimedia learning. *Applied Cognitive Psychology, 29*(5), 702–712. doi: 10.1002/acp.3154

Bates, D., Mächler, 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

Bezemer, J., & Kress, G. (2008). Writing in multimodal texts: A social semiotic account of designs for learning. *Written Communication, 25*(2), 166–195. doi:10.1177/0741088307313177

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.

Castelhano, M. S., & Rayner, K. (2008). Eye movements during reading, visual search, and scene perception: An overview. In K. Rayner, D. Shen, X. Bai, & G. Yan (Eds) *Cognitive and cultural influences on eye movements* (pp. 3–33). Tianjin: Tianjin People's Publishing House.

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. doi: 10.12691/education-3-9-5

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., & 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

Chen, S. C., Hsiao, M. S., & She, H. C. (2015). The effects of static versus dynamic 3D representations on 10th grade students' atomic orbital mental model construction: Evidence from eye movement behaviors. *Computers in Human Behavior, 53,* 169–180. doi: 10.1016/j.chb.2015.07.003

Conklin, K., Alotaibi, S., Pellicer Sanchez, A., & Vilkaitė-Lozdienė, L. (2020). What eye-tracking tells us about reading-only and reading-while-listening in a first and second language. *Second Language Research, 36*(3), 257–276. Doi: 10.1177/0267658320921496

Daly, A., & Unsworth, L. (2011). Analysis and comprehension of multimodal texts. *Australian Journal of Language and Literacy, 34*(1), 61–80.

Diao, Y., & Sweller, J. (2007). Redundancy in foreign language reading comprehension instruction: Concurrent written and spoken presentations. *Learning and Instruction, 17,* 78–88. doi: 10.1016/j.learninstruc.2006.11.007

Farias, M., Gregersen, T., Obilinovic, K., & Orrego, R. (2014). Evaluating types and combinations of multimodal presentations in the retention and transfer of concrete vocabulary in EFL learning. *Signos, 47*(84), 21–39. doi: 10.4067/S0718-09342014000100002

Han, Z., & Chen, C. A. (2010). Repeated-reading-based instructional strategy and vocabulary acquisition: A case study of a heritage speaker of Chinese. *Reading in a Foreign Language, 22*(2), 242–262.

Hannus M., & Hyönä, J. (1999). Utilization of Illustrations during Learning of Science Textbook Passages among Low- and High-Ability Children. *Contemporary Educational Psychology, 24*(2), 95–123. doi:10.1006/ceps.1998.0987

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32,* 717–742. doi: 10.1006/jmla.1993.1036

Holsanova, J. (2014). Reception of multimodality: Applying eye tracking methodology in multimodal research. In C. Jewitt (Ed.), *The Routledge Handbook of Multimodal Analysis* (pp. 287–298). New York: Routledge.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3), 346–363. doi: 10.1002/bimj.200810425

Jewitt, C., Bezemer, J., & O'Halloran (2016). *Introducing Multimodality.* New York: Routledge.

Jewitt, C., & Kress, G. (2003). *Multimodal literacy.* New York: Peter Lang.

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., & 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: Cambridge University Press.

Kress, G. (2000). *Multimodality: A Social Semiotic Approach to Communication*. London: Routledge.

Kress, G. & Van Leeuwen, T. (1996). *Reading images: The grammar of visual design.* London: Routledge.

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

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

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

Mayer, R. (2014a). Introduction to Multimedia Learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 1–24). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139547369.002

Mayer, R. (2014b). Cognitive Theory of Multimedia Learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 43–71). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139547369.005

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

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. doi: 10.1037/0022-0663.90.2.312

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

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91,* 358–368. doi: 10.1037/0022-0663.91.2.358

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

Omaggio, A. C. (1979). Pictures and second language comprehension: Do they help? *Foreign Language Annals, 12,* 107–116.

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 Sanchez, A., Tragant, E., Conklin, K., Rodgers, M., Serrano, R., & Llanes, A. (2020). Young learners' processing of multimodal input and its impact on reading comprehension: An eye-tracking study. *Studies in Second Language Acquisition.* Available First view: doi: 10.1017/S0272263120000091

Plass, J.L., Chun, D.M., Mayer, R.E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second language multimedia learning environment. *Journal of Educational Psychology, 90*, 25–36. doi: 10.1037/0022-0663.90.1.25

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: Cambridge University Press. doi:10.1017/CBO9780511816819.030

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. doi: 10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62,* 1457–1506. doi: 10.1080/17470210902816461

Rosa, E. C., & Arnold, J. E. (2017). Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language, 94,* 43–60. doi: 10.1016/j.jml.2016.07.007

Royce, T. (2002). Multimodality in the TESOL classroom: Exploring visual-verbal synergy. *TESOL Quarterly, 36*(2), 191–205.

Scheiter, K., & Eitel, A. (2017). The use of eye tracking as a research and instructional tool in multimedia learning. In C. Was, F. Sansost, & B. Morris (Eds.). *Eye-tracking technology applications in educational research* (pp. 143–165). Hershey PA, USA: Information Science Reference.

Schmidt-Weigand, F. (2011). Does animation amplify the modality effect–or is there any modality effect at all? *Zeitschrift für Padagogische Psychologie, 25*(4), 245–256. doi: 10.1024/1010-0652/a000048

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.

Schnotz, W., & Baadte, C. (2008). Domain learning versus language learning with multimedia. In M. Farías & K. Obilinovic (Eds.), *Aprendizaje multimodal/Multimodal learning* (pp. 21-49). Santiago de Chile: Publifahu USACH.

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,* Available first view: https://doi.org/10.1515/applirev-2018-0102

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*, 70–96.

Takacs, Z. K., & Bus, A. G. (2016). Benefits of motion in animated storybooks for Children's visual attention and story comprehension. An eye-tracking study. *Frontiers in Psychology, 7,* 1–12. doi: 10.3389/fpsyg.2016.01591

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

Unsworth, L. (2014). Multimodal reading comprehension: curriculum expectations and large-scale literacy testing practices, *Pedagogies: An International Journal, 9*(1), 26–44. doi: 10.1080/1554480X.2014.878968

Unsworth, L., Thomas, A., & Bush, R. (2004). The role of images and image-text relations in group 'basic skills tests' of literacy for children in the primary school years. *Australian Journal of Language and Literacy, 27(*1), 46–65.

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* (6), 667–686.

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.