Learning and Instruction 71 (2021) 101397

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

# Effects of picture-word integration on reading visual narratives in L1 and L2

Yen Na Yum [a,b,*], Neil Cohn [c], Way Kwok-Wai Lau [a,b]

[a] Department of Special Education and Counselling, The Education University of Hong Kong, 10 Lo Ping Road, NT, Hong Kong
[b] Integrated Center for Wellbeing, The Education University of Hong Kong, Hong Kong
[c] Tilburg Center for Cognition and Communication, Tilburg University, P.O. Box 90153, 5000, LE, Tilburg, the Netherlands

ARTICLE INFO

*Keywords:*
Bilingual
Multimodal reading
Perceived ease
Situational interest
Visual narratives

ABSTRACT

Multimodal education materials are pervasive in language learning. This study investigated the causal mechanisms of multimodal reading effects in first language (L1) and second language (L2). Seventy-five adult bilingual readers in Hong Kong read Chinese and English passages with different degrees of picture-word integration in a within-subject design. Results showed that tight text-picture integration facilitated better comprehension than independent text-picture presentation in L2, but not L1. Perceived ease and interest differentially mediated multimodal reading performance for L1 and L2 passages. Importantly, separate images in L2 passages led to poorer comprehension accuracy relative to plain text, but tended to have higher ratings of ease and interest, indicating that readers may be overconfident in their multimodal reading performance. In general, results support the notion that integration of text and pictures can moderate the process of meaning making, and these may differ depending on the language presented to a bilingual reader.

## 1. Introduction

Multimodal education materials are pervasive from picture books in early childhood education to lecture slides in university classrooms. The wide usage is based on the *multimedia learning effect* — information presented by words and pictures facilitate learning and memory relative to information presented by words alone (for review, see Mayer, 2009). Among different types of multimodal text, visual narratives including comic strips and graphic novels have gathered recent research interest. Visual narratives present information in a sequence of images, which typically convey events or situations, and often combine multimodally with written text (Cohn & Magliano, 2020). Emerging research on visual narratives has found that similar brain responses appear to the processing of visual narrative sequences and sentences (for review, see Cohn, 2020b), including similar patterns of impairment (Coderre et al., 2018), implying overlapping cognitive mechanisms. These findings imply a tighter relationship between language and visual-graphic communication than has previously been acknowledged, and also suggest unique processing requirements for multimodal text (Cohn, 2013).

Much of the existing research on the educational use of visual narratives examined whether they can enhance reading comprehension. According to the PISA reading test framework (OECD, 2019), reading literacy is the readers' ability to understand, use, evaluate, reflect on and engage with text to achieve their purposes. These abilities are built on three broad categories of reading skills: access and retrieve; integrate and interpret; and reflect and evaluate. Several classroom-based studies have shown that non-native readers of low second language (L2) proficiency benefitted from reading comics compared to reading plain text in L2, even though the text contents were identical (Liu, 2004; Merc, 2013). A recent experimental study explored how graphic novels affected L2 English comprehension of psychology concepts among university students (Wong, Miao, Cheng, & Yip, 2017). Participants had better reading comprehension performance when materials were presented as a visual narrative than as plain text, regardless of English abilities. While a growing body of literature is investigating the use of comics for L2 learning in classroom or experimental settings, few studies have directly compared multimodal effects on reading comprehension in first language (L1) and L2 reading. The current study filled this gap by using a within-subject design in adult bilingual participants to systematically assess the role of presentation language in multimodal comprehension.

### 1.1. Integration of text and image in visual narratives

An important trait that characterizes visual narratives is the tight integration of text and image in their graphical layouts. Cohn (2013)

* Corresponding author. Department of Special Education and Counselling, The Education University of Hong Kong, 10 Lo Ping Road, NT, Hong Kong.
*E-mail addresses:* yyum@eduhk.hk (Y.N. Yum), neilcohn@visuallanguagelab.com (N. Cohn), waylau@eduhk.hk (W.K.-W. Lau).

https://doi.org/10.1016/j.learninstruc.2020.101397
Received 13 March 2020; Received in revised form 12 August 2020; Accepted 15 August 2020

argued that written language and pictures relate across four levels of interfaces (Fig. 1). An *inherent* relation is when text appears within the world of the images. An *emergent* relation occurs using visual "carriers" of text, where the "tail" indexically links the carrier (like a thought bubble) to its visual "root" (like a thinker). This relationship tightly joins the text to the image through a conventionalized interface. An *adjoined* relation uses disconnected captions that float above the image content. Finally, an *independent* relation keeps text and image completely separate, such as a picture illustrating ideas of a separated caption or body text (e.g., "see Fig. 1").

Of most interest to us here is the difference between independent relations and emergent/adjoined relationships. These latter interfaces use particular "bundling" techniques, such as linking carriers and text via their tails and grouping text and picture via the frame of a panel border. These techniques employ Gestalt constraints of encapsulation and connectedness (Palmer, 1992; Palmer & Rock, 1994), which create cohesive integrated multimodal text-image units. Such integrated units are thus contrasted from the disconnected independent relations which involve no such unitization of the multimodal message, and which are more typical of educational materials like traditional textbooks. According to the idea that text and images originate in a common multimodal cognitive architecture (Cohn, 2016), the degree of the integration across text and image should predict effectiveness of communication over and above the independent features of text or image.

These types of integration methods operationalize the structural properties of various types of spatial contiguity (see Mayer, 2009) observed between text and image. Theoretically, greater space between text and images has been posited as contributing to greater extraneous cognitive load because it creates a split attention across the overall message (Sweller, 2005). Research on this in the context of instruction shows that separating text and image consistently leads to worse retention and transfer performance. A meta-analysis of 37 studies found a large mean weighted effect size ($d = 0.72$) for the spatial contiguity effect (Ginns, 2006). This effect was stable across individual or group tests, static image or animation presentation, and primary students to college-level adults. All studies reviewed in this meta-analysis were in science learning, where images usually label new terms or processes, and spatial proximity is the main method to link the independently-presented image and text. A distinction may be drawn where, in addition to spatial distance, visual narratives use conventional symbols, such as tails and panels, for cue-based text-image integration.

Several theoretical frameworks attributed the multimedia learning benefits to separate systems that support verbal versus graphic information, such as the dual-coding theory (Paivio, 1991) and the cognitive theory of multimedia learning (Mayer, 1997). These theories posited that text and image are separately encoded and recalled by verbal memory and spatial memory subsystems. In L2 reading, the verbal memory subsystem would be challenged, but the spatial memory subsystem would be unaffected. This would predict main effects of presentation mode and presentation language on reading performance, but no interaction. Other theoretical accounts treat verbal and pictorial modalities as parts of a single broader multimodal communication system, such as the multimodal parallel architecture (Cohn, 2016) and the integrated model of text and picture comprehension (Schnotz, 2005; Schnotz & Bannert, 2003). While text and images are clearly different representationally, comprehension across modalities draw on overlapping semantic memory resources and are integrated into a single mental model (Coderre et al., 2018; Kutas & Federmeier, 2011; Magliano, Larson, Higgs, & Loschky, 2016). These unitary frameworks would predict that multimedia learning benefits are strong in L2 reading because images could supplement the textual information and form an enriched mental model. At the same time, readers would be more sensitive to text-image integration effects in L2 compared to L1, assuming that presenting passages in L2 increases the intrinsic cognitive load of the materials (Sweller, 2005). Since the working memory system supports both textual and graphic information processing, there is less working memory capacity to process the extraneous cognitive load caused by instructional design (e.g., degree of text-image integration) of materials presented in L2.

### 1.2. Mechanisms of multimodal reading effects

While there are several theoretical accounts of multimodal reading, the mechanisms that enable combined text and visual representations to facilitate reading comprehension remain unclear. Here we focused on two factors that may causally mediate multimodal effects, namely

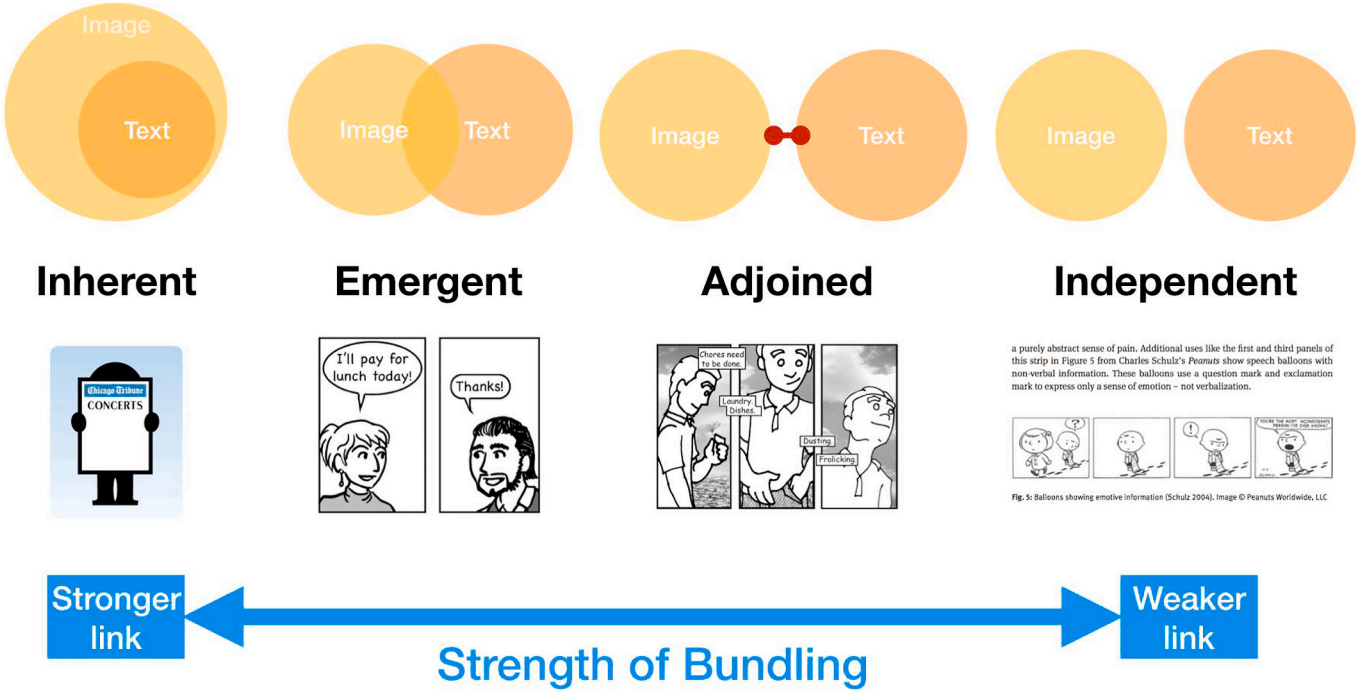


**Fig. 1.** Four types of text-image relationships with varying degrees of integration.

situational interest and perceived ease. Situational interest refers to the temporary arousal of attention and motivation to learn about a presented topic, task, or object based on its specific features (Schiefele, 2009). Higher level of situational interest of text typically correlates with longer reading time, possibly reflecting increased cognitive engagement (Leutner, 2014). Enhanced interest may also have other behavioral manifestations, for instance, after reading about psychology concepts in visual narrative form, students were motivated to select more references for future reading (Wong et al., 2017). Magner, Schwonke, Aleven, Popescu, and Renkl (2014) found that decorative pictures fostered the feeling component of situational interest (ratings of how entertaining the passage is), while the value component of interest (ratings of how useful the passage is) was unaffected. The increase in interest improved far transfer of learning but had no effect on near transfer. Thus, their results showed that situational interest mediated multimodal effects of more cognitively demanding tasks by promoting learning.

Perceived ease is a metacognitive measure, reflecting confidence in comprehension and prediction of accurate responses. Previous studies have shown that visual narratives facilitated reading comprehension among students with low prior knowledge of reading contents, which Eitel, Scheiter, and Schüler (2013) explained by an increase in reading ease due to inclusion of graphics. In experimental context where learning materials were counterbalanced, it could also be construed as extraneous cognitive load, defined as the cognitive processing induced by instructional design independent from the difficulty of the learning materials (Sweller, 2005). The mental load that is invested by the learner is argued to be reliably and accurately monitored based on subjective ratings (Sweller, Van Merrienboer, & Paas, 1998). Other reports found that while multimodal presentation boosted readers' confidence in learning, it did not improve comprehension performance (Lindner, Eitel, Barenthien, & Köller, 2018; Serra & Dunlosky, 2010). The inflation of judgment of ease, known as the multimedia heuristic, may create over-confidence that reduces cognitive engagement and is undesirable for learning and retention (Serra & Dunlosky, 2010).

Both interest and ease have been used as mediators and outcomes in multimodal reading, but little direct evidence differentiate between their potential mediating effects. One reason to tease apart interest and ease is that they may exert opposite effects on reading time and effectively cancel each other out. While both interest and ease may be expected to increase comprehension performance, there were inconsistent reports regarding improvement to actual performance (e.g., Lindner et al., 2018; Magner et al., 2014). In other words, changes in interest and ease may be by-products of multimodal reading, without mediating changes in reading outcomes.

### 1.3. The current study

This study investigated reading time and comprehension performance in Chinese and English passage reading in three levels of text-image integration. Individual differences in L1 and L2 proficiency and visual language fluency were used as covariates. Ratings of interest and reading ease were collected for each passage and evaluated as both outcomes and mediators of reading performance. The study aimed to address two broad research questions.

#### 1.3.1. Do reading outcomes in L1 and L2 interact with different levels of text-picture integration?

Based on the literature, it was expected that greater facilitation would be seen in texts with emergent/adjoined relations than texts with independent relations (Cohn, 2013). We also predicted that L2 readers would benefit more from added images, which act as a scaffold for construction of a mental model. Thus, we hypothesized that while readers may show lower reading comprehension performance in L2 relative to L1, they would show greater multimodal facilitation when reading in their L2. Facilitation of tight text-picture integration was predicted to be stronger in L2 reading, since effects of extraneous cognitive load tended to manifest when intrinsic cognitive load is high.

#### 1.3.2. Do interest or ease mediate multimodal reading in L1 and L2?

It was predicted that both texts with separate images or integrated images would increase ratings of interest and perceived ease (Lindner et al., 2018; Wong et al., 2017). High situational interest was hypothesized to prolong reading time, while high perceived ease would shorten reading time. The mediating effects of interest and ease on comprehension accuracy, if any, may be stronger in L2 than L1, as task difficulty may moderate the mediation (Magner et al., 2014).

## 2. Method

### 2.1. Participants

A power analysis showed that 24–38 participants were needed to achieve 0.8 power to reveal medium effect sizes ($d = 0.39$) in a counterbalanced design with one fixed factor and two crossed random effects (Judd, Westfall, & Kenny, 2017). Since we examined two fixed effects with interaction, the target sample size was doubled. With reference to mixed effects mediation analysis, the minimum sample size for 0.8 power was 44–82 participants for subject intraclass correlation (ICC) = 0.1–0.9, again assuming effect sizes of $d = 0.39$ (Pan, Liu, Miao, & Yuan, 2018). We recruited eighty university-educated adults who gave written informed consent and received supermarket coupons for taking part in the experiment. According to the total reading times from all conditions, outlying data from 2 participants were excluded based on the probability of Mahalanobis distance analysis. An additional 3 participants were excluded based on self-reported learning or sensory disabilities, yielding 75 participants for data analyses (60 females; mean age: 22.1 years, range: 18–39 years). Language background of the participants are summarized in Table 1. All participants were native speakers of Cantonese Chinese (henceforth Chinese) who learned English as L2 in school and had completed public examination for university entrance in Hong Kong. Participants were dominant in Chinese in terms of amount of use and self-rated proficiency.

### 2.2. Procedure

After completing the language background questionnaire, participants read six short passages on a computer monitor individually in a quiet room. Participants were randomly assigned to one of twelve lists with alternating languages and modes, in which the conditions of the passages were counterbalanced across participants. Pages were presented one at a time and self-paced to simulate natural reading. When participants finished reading a page, they pressed a button for the next page, but could not re-visit previous pages. After each passage, participants answered two multiple choice comprehension questions, then provided three ratings on whether the passage was easy, interesting, or familiar. Presentation of stimuli and recording of reading times and responses to comprehension questions were controlled by EPrime2.0 (Psychology Software Tools). The study protocol has received approval

**Table 1**
Language background of participants, with mean values (SDs).

| | Chinese | English |
|---|---|---|
| Public examination grades[a] | 3.71 (0.82) | 3.65 (0.76) |
| Amount of daily language use | 75.6% (16.0) | 18.5% (14.5) |
| Oral production level[b] | 8.17 (1.47) | 6.05 (1.55) |
| Oral comprehension level[b] | 7.67 (1.36) | 5.97 (1.38) |
| Reading comprehension level[b] | 7.53 (1.46) | 5.93 (1.36) |

*Notes*.
[a] Public examination grades could range from 1 to 7, with greater values indicating better grades.
[b] Self-rated on a scale of 1–10 with 10 being the highest proficiency.

from the human research ethics committee at a university in Hong Kong.

### 2.3. Stimuli and measures

#### 2.3.1. Passages

The reading comprehension task consisted of six short passages with non-fiction topics (biology, economics, history, medicine, philosophy, and physics), which were presented in plain text, text with separate images, and text with integrated images (see Fig. 2 for sample stimuli). Each passage had Chinese and English versions which were standardized to a length of four pages. The passages were expository in nature, written for general interest, and were of moderate difficulty. The English passages had a Flesch reading ease of 59.0 ± 10.7 (Flesch, 1948) and mean lexical density of 54.6% ± 3.9 (proportion of content words over all words). Texts were controlled for the mean numbers of words (Chinese = 318.7 ± 10.9; English = 343.2 ± 5.2) and the mean number of characters without space (Chinese = 511 ± 20; English = 1592 ± 115) within each language. The mean word length of Chinese words was 1.62 ± 0.09 characters, while mean word length of English words was 5.41 ± 0.34 letters across the passages. The mean lexical frequency of the Chinese words was 7212 ± 1895 per million of words according to SUBTLEX-CH, a database formed from film subtitles (Cai & Brysbaert, 2010) and frequency of the English words was 5659 ± 0.22 (SUBTLEX-US; Brysbaert & New, 2009). In each Chinese and English passage, between eight to ten words which were technical terms (e.g., "neurons") or proper names (e.g., "Turing") were not found in the databases. Across the three presentation modes, the texts presented were identical. For passages presented in plain text, text was presented in the center of the screen. Passages presented in integrated images mode were adapted from existing comics, which were standardized into the same size and presented in greyscale. The numbers of image panels in the passages were matched (13.5 ± 2.6). For passages in separate images mode, words were presented in the center of the screen in the same position as plain text mode, while image panels without text were presented in the upper and lower portions of the screen.

#### 2.3.2. Comprehension questions

Multiple choice questions were used to assess factual understanding and recall of the passage contents, i.e., ability to access and retrieve information for comprehension, without requiring transfer knowledge. For example, a passage contained the sentence "The tetanus bacteria *clostridium tetani* is widely distributed as spores in soil and the feces of many animals." A comprehension question was "What form does the tetanus bacteria usually take?" and the answer was "spores in soil." The accuracy measure was based on the responses to two comprehension questions per passage (Cronbach's α = 0.54; mean inter-item correlation = 0.383). While the reliability indices were low because of the low number of questions, the inter-item correlation of the questions showed internal consistency and a suitable degree of item overlap (Piedmont, 2014, pp. 3303–3304). Each question had four answer choices, so chance level was 25%. Accuracy was scored in binary mode, either correct or incorrect, with no score penalty on incorrect answers.

#### 2.3.3. Language proficiency

Participants' language proficiency levels were measured by subject grades for Chinese and English language in the Hong Kong Diploma of Secondary Education Examination. This was the public examination taken by secondary school graduates for university entrance in Hong Kong, administered by the Hong Kong Examinations and Assessment Authority (HKEAA). The examination used standards-referenced reporting with annual calibration exercises to explicit and fixed performance standards, ensuring that scores across years would reflect same levels of performance (Hong Kong Examinations and Assessment Authority, 2018). While the grades showed similar numerical values for Chinese and English on the 1–7 scale (see Table 1), the examinations evaluated native and non-native language abilities and were not assumed to be directly comparable. For reference, the participants' average grade of 3.65 on the English subject test is roughly equivalent to an overall band score of 6 in the International English Language Testing System (IELTS) and corresponds to B2 level in the Common European Framework Reference (CEFR) (Hong Kong Examinations and Assessment Authority, 2013); (International English Language Test System, 2020).

#### 2.3.4. Visual language fluency

The Visual Language Fluency Index (Cohn, 2020a) measured the frequency and expertise of reading and drawing of graphical materials in past and present (8 items, Cronbach's α = 0.80). The mean score of VLFI

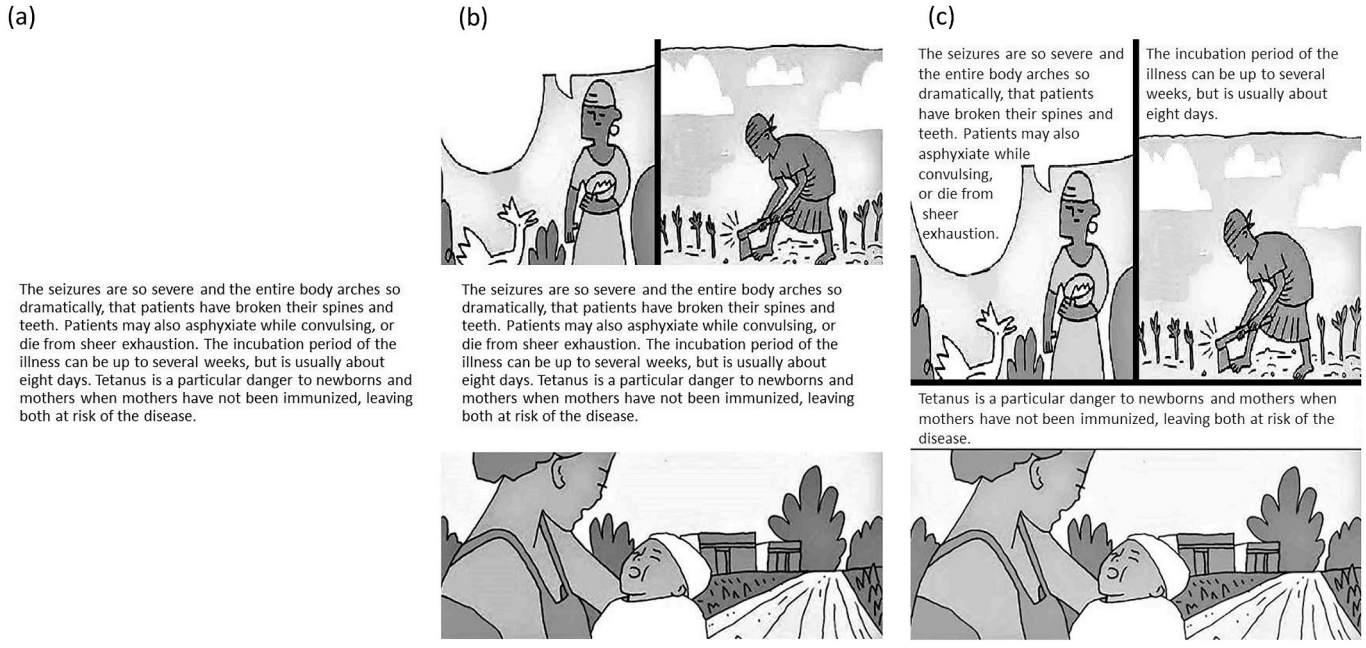


**Fig. 2.** Sample page in an English passage in (a) plain text, (b) separate images, and (c) integrated images modes. Adapted from "Stopping Tetanus in Mothers and Newborns", Art of Saving a Life Project, artwork by Darryl Cunningham.

for all participants was 7.8 ± 7.5, considered an overall low fluency in visual language (<8: low, ~12: average, >22: high; Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012).

#### 2.3.5. Interest, ease, and familiarity ratings

For each passage, participants provided three ratings by 5-point Likert scales (1 = strongly disagree, 5 = strongly agree). Situational interest was measured by agreement with "This reading was interesting" (Magner et al., 2014). Perceived ease was measured by rating of "This reading was easy to understand," a positively phrased measure of extraneous cognitive load (Sweller et al., 1998). Prior knowledge of the passage was measured by rating of "I am familiar with the content of this reading."

### 2.4. Data analyses

Two sets of analyses were run to address the research questions—linear mixed-effects regression models (Baayen, Davidson, & Bates, 2008) and mixed-effects causal mediation analyses (Imai, Keele, Tingley, & Yamamoto, 2011). Mixed-effects models account for random effects that occur from drawing a sample from a population, by adjusting the random intercepts and random slopes of the fitted lines. The models were computed on the statistical platform R (R core team, 2017) using the Satterthwaite method to approximate denominator degrees of freedom using the packages *lme4* (Bates, Mächler, Bolker, & Walker, 2015), *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017), and *ordinal* (Christensen, 2019).

#### 2.4.1. Mixed-effects regression models

A total of four models were fitted for four dependent variables: situational interest and perceived ease were modelled by ordinal mixed models, reading time of each page was modelled by a linear mixed effects models, while the logged odds for correct response over incorrect response was modelled by a generalized linear mixed-effects model with a binomial distribution (logit linking function). The main independent variables were presentation language (Chinese and English) and presentation mode (plain text, separate images, and integrated images) and their interactions. Presentation language was examined using a sum contrast, where L1 Chinese was compared to the mean of Chinese and English. The three levels of presentation mode were examined using a Helmert contrast. First, results of plain text were compared to the average of text with integrated or separate images for the effect of multimodal presentation (i.e., reading with or without images). Then, results of text with integrated or separate images were compared to each other for the effect of degree of text-image integration. Language grades, visual language fluency, and topic familiarity were evaluated as covariates. Language grades were language-specific, i.e., Chinese grades were used to estimate effects of passages presented in Chinese and English grades were used to estimate effects of passages presented in English. In the reading time model, page order within the passage was included as a covariate to control for potential reading time differences induced by meaning integration across pages. Longer reading time at the beginning of a passage may be due to lack of context, while longer reading time towards the end of a passage may be due to wrap-up effects where participants integrate information across pages.

The $\beta$ estimates, confidence intervals, and $p$ values of planned contrasts and covariates were reported. For significant interactions, follow-up pairwise comparisons of conditions were reported with Bonferroni corrections of the $p$ values. In all models, $z$-score transformation was done for continuous predictors (language grades, interest, ease, familiarity, page number, and total reading time). The centering of the predictors reduced collinearity and the standardization facilitated comparison of effects for predictors on different scales. Initial models included fixed factors, i.e., independent variables and covariates, and random intercepts of participant and stimulus (passage, page, or comprehension question). The random intercept of participant captures the individual differences in ratings, reading speed, and accuracy, while random intercept of stimulus capture the different processing difficulty of the stimulus. Where applicable, these were reported with the intra-class correlations (ICCs) which reflect the proportion of variance explained by the grouping structure (i.e., participant or stimulus). Then random slopes for significant fixed effects were fitted to capture the possibility that the effects differ across different participants or stimulus. The random slopes were included only if the corresponding likelihood ratio tests were significant, suggesting model improvement (Pinheiro & Bates, 2006).

#### 2.4.2. Causal mediation of interest and ease

Fixed effects parameter estimates and 95% confidence intervals were calculated using the bootstrap method based on 3000 bootstrap samples with replacement. The analyses were done with the *mediation* package (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014). Presentation mode was the main independent variable and the reading performance measures of reading time and comprehension accuracy were dependent variables. Situational interest and perceived ease were fitted as mediators. Presentation modes were modelled in pairwise manner, i.e., effects of separate images and integrated images were compared to the control condition of plain text. Since presentation mode and language did not have significant interaction effects on perceived ease or interest in the mixed effects models, presentation languages (L1 and L2) were modelled separately for ease of interpretation. Language grades, visual language fluency, and familiarity were control variables in the models. Since the statistical methods to examine random effects from different sources were still in development, only random intercepts for participants were fitted because participants' individual differences were greater compared to heterogeneity of the passages.

## 3. Results

Descriptive statistics of the dependent variables (interest, ease, reading time, and comprehension accuracy) are summarized in Table 2, while the correlations among all measured variables separated by condition are presented in the Appendix.

### 3.1. Mixed effects regression models

#### 3.1.1. Interest

The final model for situational interest included random intercepts for participant and passage only, since random slopes were not significant. The parameter estimate results for the model are in Table 3. As can be seen in Table 3, the multimodal contrast of presentation mode was marginally significant, showing a trend for multimodal passages to be judged as more interesting than plain text. Meanwhile, the integration contrast indicated that passages with separate images were rated as more interesting than passages with integrated images. Passages presented in L2 English received significantly higher interest ratings than passages presented in L1 Chinese. The interaction of mode and language did not reach significance for either contrast. Higher visual language fluency and higher topic familiarity significantly increased interest ratings.

#### 3.1.2. Ease

The final model for perceived ease included random intercepts for participant and passage only, since models with random slopes did not converge. As predicted, multimodal passages were judged to be easier than plain text, while the text-image integration contrast was not significant (Table 3). Passages presented in L2 English were judged to be easier than passages presented in L1 Chinese. The interaction of mode and language did not reach significance for either the multimodal or the integration contrast. Language grades had a significant effect on perceived ease, such that higher grades predicted higher ease ratings. Higher topic familiarity also led to higher perceived ease.

**Table 2**
Observed means (and standard deviations) for dependent variables as a function of language and mode of presentation. Separate = text with separate images; Integrated = text with integrated images.

| | Chinese passage | | | English passage | | |
|---|---|---|---|---|---|---|
| | Plain Text | Separate | Integrated | Plain Text | Separate | Integrated |
| Interest[a] | 2.87 (1.07) | 3.04 (0.98) | 2.91 (0.9) | 3.07 (0.99) | 3.25 (0.99) | 2.95 (0.98) |
| **Ease**[b] | 2.81 (1.06) | 2.99 (0.86) | 3.20 (0.99) | 3.16 (0.81) | 3.17 (0.92) | 3.16 (1.08) |
| **Reading time**[c] | 21.76 (13.87) | 21.94 (13.22) | 24.65 (14.83) | 30.93 (15.29) | 33.90 (16.89) | 35.67 (15.37) |
| **Accuracy**[d] | 67% (0.47) | 61% (0.49) | 61% (0.49) | 71% (0.46) | 55% (0.50) | 67% (0.47) |

Notes.
[a] Higher values indicate greater interest in the reading.
[b] Higher values represent higher perceived ease.
[c] Reported in seconds per page.
[d] Mean accuracy of two comprehension questions.

#### 3.1.3. Reading time

Reading time data were trimmed such that pages that were viewed for less than 2 s (1.0%) or were different from the overall mean for more than three standard deviations (1.3%) were excluded in the analyses. The final model for reading time per page included random intercepts for participant and page, and random slopes for presentation mode and language by participant. The ICC for subject random effect was 0.536 and that for page was .229. The parameter estimate results for the reading time model are in Table 4.

Both the multimodal and the integration main effect contrasts were significant, such that reading time was significantly longer for multimodal passages than for plain text, and passages with integrated images were read for significantly longer than passages with separate images (Table 4). Passages presented in English were read for longer than passages presented in Chinese. The interaction of mode and language was significant only for the multimodal contrast (Table 4 and Fig. 3). Participants showed a clear multimodal effect in L2 passages, where passages with separate and integrated images were both read for longer than plain text ($t = 5.76$, $p < .001$ and $t = 3.97$, $p = .001$, respectively); in L1 passages, participants had longer reading time for integrated mode relative to plain text ($t = 4.15$, $p < .001$), but not for separate mode relative to plain text ($t = 0.18$, $p = 1.00$). The text-image integration contrast did not interact with presentation language. Reading time decreased as page number increased, suggesting that participants sped up towards the end of a passage.

#### 3.1.4. Comprehension accuracy

The final model for comprehension accuracy included random intercepts for participant and passage but not random slopes, since models with random slopes did not converge. The ICC for subject random effect was 0.204 and that for question was 0.136 on logistic scale. The parameter estimate results for the model of comprehension question accuracy are in Table 5.

The multimodal contrast was not significant in this model, while the integration contrast was marginally significant with integrated mode being responded to more accurately than the separate mode (Table 5). Main effect of presentation language on comprehension accuracy was not significant. The multimodal contrast did not interact with presentation language, while the integration contrast was significantly different depending on presentation language (Table 5 and Fig. 4). Passages with integrated images elicited higher accuracy than passages with separate images in L2 passages ($z = -3.07$, $p = .032$) but not L1 passages ($z = 0.37$, $p = 1.00$). Higher visual language fluency showed a trend of decreasing comprehension accuracy, while higher topic familiarity significantly increased comprehension accuracy.

### 3.2. Causal mediation analyses

Patterns of results for the causal mediation analyses were summarized in Table 6 and the causal pathways were illustrated in Fig. 5.

#### 3.2.1. Reading time

As predicted, when reading in L1, inclusion of separate images increased situational interest, which increased reading time. However, the mediating effect of interest in L1 passage with integrated images did not reach significance. For perceived ease, both separate and integrated modes elicited higher ease judgments relative to plain text, which in turn decreased reading time. Direct effects of integrated images were significant while those of separate images were not.

When reading in L2 English, participants were more interested in passages with separate images, which led to longer reading time. Interest did not mediate effects of passages with integrated images. Perceived ease was higher when integrated images were included, but this did not significantly change reading time. Significant direct effects of images were found in both separate and integrated image presentation modes.

#### 3.2.2. Comprehension accuracy

For L1 passages, interest did not mediate effects on comprehension of text with either separate or integrated images. Meanwhile, higher perceived ease in texts with separate and integrated images mediated the mode effects, leading to better comprehension. Direct effects of presentation mode on comprehension accuracy of L1 passages were insignificant for either separate or integrated modes.

In contrast, neither interest nor ease mediated multimodal effects of L2 passages with separate or integrated images on comprehension accuracy. Separate images led to poorer comprehension relative to plain text, as shown by the negative direct effects on comprehension accuracy.

## 4. Discussion

This study examined passage reading in bilingual adults to address two research questions on effects of presentation modes and languages, and mediation effects of interest and ease on presentation modes. Regarding the first question, we found that a significant multimodal effect for reading time but not comprehension accuracy, such that passages with pictures were read for a longer time but were not responded to more accurately. Additionally, both reading time and comprehension accuracy showed text-image integration effects. Longer reading time for integrated mode over separate mode was found, with a significant main effect for both presentation languages. Confirming our predictions, tight integration of text and image facilitated reading comprehension accuracy to a greater extent than images independently presented with text, but only in L2. Text with separate images may actually interfere with L2 comprehension, as discussed further below. This supported functional distinction of the categories of text-picture interface proposed by Cohn (2013) and the overall hypothesis for unified processing for text and image sequence.

The current study used narrative text with different topics to examine more general effects of multimodal reading. The contents of the passages were not particularly difficult conceptually, but were more

**Table 3**
Parameter estimate results of mixed-effects ordinal logistic regression models for situational interest and perceived ease.

Situational interest (1–5 ratings)

**Model equation:** Interest ~ Mode x Presentation language + Familiarity + Language grade + Visual language fluency + (1 | Participant) + (1 | Passage)

**Model fit:** log Likelihood = −554.74, AIC = 1137.47

| **Fixed Effects** | ***β*** | ***SE*** | **95% CI** | | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| Multimodal contrast | −0.13 | 0.06 | [ −0.25 | 0.00 ] | −1.94 | .052 |
| Integration contrast | 0.32 | 0.11 | [ 0.45 | 0.08 ] | 2.94 | .003* |
| Presentation language | −0.19 | 0.09 | [ −0.37 | −0.02 ] | −2.14 | .032* |
| Visual language fluency | 0.05 | 0.02 | [ 0.01 | 0.08 ] | 2.53 | .011* |
| Language grade | 0.12 | 0.14 | [ −0.15 | 0.39 ] | 0.91 | .365 |
| Familiarity of the topic | 0.93 | 0.11 | [ 0.72 | 1.14 ] | 8.72 | <.001* |
| Multimodal contrast x Presentation language | 0.01 | 0.16 | [ −0.13 | 0.12 ] | 0.03 | .973 |
| Integration contrast x Presentation language | 0.07 | 0.19 | [ −0.17 | 0.26 ] | 0.39 | .697 |

| **Random Effects** | | | **Variance** | ***SD*** |
|---|---|---|---|---|
| Participant (intercept) | | | 0.67 | 0.82 |
| Passage (intercept) | | | 0.16 | 0.40 |

**Perceived Ease (1–5 ratings)**
**Model equation:** Ease ~ Mode x Presentation language + Familiarity + Language grade + Visual language fluency + (1 | Participant) + (1 | Passage)
**Model fit:** log Likelihood = −509.11, AIC = 1046.22

| **Fixed Effects** | ***β*** | ***SE*** | **95% CI** | | ***t*** | ***p*** |
|---|---|---|---|---|---|---|
| Multimodal contrast | −0.22 | 0.07 | [ −0.35 | −0.09 ] | −3.39 | <.001* |
| Integration contrast | −0.08 | 0.11 | [ −0.30 | 0.14 ] | 0.72 | .471 |
| Presentation language | −0.27 | 0.09 | [ −0.45 | −0.08 ] | −2.86 | .004* |
| Visual language fluency | 0.01 | 0.02 | [ −0.02 | 0.05 ] | 0.77 | .441 |
| Language grade | 0.28 | 0.14 | [ 0.01 | 0.56 ] | 2.03 | .043* |
| Familiarity of the topic | 1.24 | 0.11 | [ 1.02 | 1.47 ] | 10.82 | <.001* |
| Multimodal contrast x Presentation language | −0.08 | 0.06 | [ −0.21 | 0.05 ] | −1.23 | .217 |
| Integration contrast x Presentation language | 0.02 | 0.11 | [ −0.20 | 0.24 ] | 0.18 | .857 |

| **Random Effects** | **Variance** | ***SD*** |
|---|---|---|
| Participant (intercept) | 0.55 | 0.74 |
| Passage (intercept) | 0.50 | 0.70 |

*Notes.* Multimodal contrasts showed effects of plain text subtracting the average effects of integrated and separate modes, while the integration contrasts showed effects of separate mode subtracting effects of integrated mode. The presentation language contrasts showed estimate of L1 effects subtracting the average effects of L1 and L2.

typical of reading comprehension text in second language acquisition, as in Liu (2004) and Merc (2013). In these studies of visual narratives, the presentation language was L2, and text difficulty or participants' L2 proficiency was manipulated for comparison. In contrast, the vast majority of studies of the multimedia effect used stimuli that were adapted from diagrams and explanations of complex science concepts (e.g., Hannus & Hyönä, 1999; Lindner et al., 2018; Schnotz & Wagner, 2018). Ginns (2006) reported that the greatest spatial contiguity effect was in novice learners acquiring complex materials, where text and image contribute different information to the mental model of an unfamiliar concept.

So, one interpretation was that L1 narrative text did not require participants to seek additional information or support from the presented images (low intrinsic cognitive load), resulting in no associated multimodal effects. This would be consistent with the reading time results for L1 passages where participants spent similar amounts of time reading plain text and text with separate images, suggesting that they did not spend much time on processing separately presented images, as in previous literature (e.g., Hannus & Hyönä, 1999; Schmidt-Weigand, Kohnert, & Glowalla, 2010). In the reading time for L2 passages, text with separate images patterned with text with integrated images, and both were viewed for longer than plain text. These results indicated a choice in using images to support L2 reading, while ignoring separately presented images in skilled L1 reading. The lack of a multimodal facilitation effect could also have been influenced by participants' comic reading expertise. Previous studies have shown that readers with lower expertise in reading comics tended to attend more to the text than the images (Kirtley, Murray, Vaughan, & Tatler, 2018; Laubrock, Hohenstein, & Kümmerer, 2018; Zhao & Mahrt, 2018), and our participants indeed had fairly low scores for visual language fluency. Thus, participants' low visual language fluency may have pushed them to focus more on the contents of the text at the expense of multimodal integration of text and images.

### 4.1. Interest and ease as mediators of multimodal effects

In line with our predictions, inclusion of separate images in L1 and L2 passages increased situational interest, which increased reading time. However, this increase in interest did not mediate comprehension accuracy as in Magner et al. (2014). One possible interpretation was that the higher interest ratings for separate mode were based on visual interest, solely because the presentation format was novel for the participants. If the images did not increase interest in the textual contents, one would expect no conceptual facilitation in comprehension accuracy. Meanwhile, passages with integrated images surprisingly led to lower interest in L1 reading (Fig. 5), perhaps due to the overall low visual language fluency. While participants still preferred to have a multimodal message over plain text (separate mode), the integrated mode may have been harder as dissecting the visual language becomes a cognitive load (as suggested by VLFI being associated with both situational interest and poor reading comprehension).

The mediation analyses showed divergent causal pathways for multimodal effects and perceived ease in L1 and L2 reading. Ease mediated multimodal effects for L1 reading, where greater perceived ease in text with images decreased reading time and increased comprehension accuracy. Thus, subjective ratings of ease for L1 passages accurately reflected readers' mental effort and objective comprehension performance. Meanwhile, mediating effects of ease were not found at all in L2 reading time or comprehension accuracy. In other words, readers felt that the integrated images made L2 comprehension

**Table 4**
Parameter estimate results of the mixed-effects linear regression model for reading time (seconds).

| Model equation: Reading time ~ Mode x Presentation language + Familiarity + Language grade + Visual language fluency + (Mode + Presentation language \| Participant) + (1 \| Page) | | | | | | |
|---|---|---|---|---|---|---|
| Model fit: $R^2$ marginal = .213, $R^2$ conditional = .694, AIC = 13,223.14 | | | | | | |
| Fixed Effects | *β* | *SE* | 95% CI | | *t* | *P* |
| Multimodal contrast | −2.85 | 0.51 | [ −3.85 | −1.86 ] | −5.63 | <.001* |
| Integration contrast | −2.46 | 0.65 | [ −3.74 | −1.18 ] | −3.77 | <.001* |
| Presentation language | −5.35 | 0.80 | [ −6.93 | −3.78 ] | −6.68 | <.001* |
| Visual language fluency | −0.78 | 1.11 | [ −2.97 | 1.40 ] | −0.70 | .484 |
| Language grade | −0.67 | 0.52 | [ −1.68 | 0.34 ] | −1.29 | .200 |
| Familiarity of the topic | −0.05 | 0.30 | [ −0.63 | 0.53 ] | −0.16 | .871 |
| Page number within the passage | −4.80 | 0.74 | [ −6.24 | −3.35 ] | −6.51 | <.001* |
| Multimodal contrast x Presentation language | 1.09 | 0.46 | [ 0.20 | 1.99 ] | 2.39 | .017* |
| Integration contrast x Presentation language | −0.79 | 0.53 | [ −1.83 | 0.26 ] | −1.47 | .141 |
| **Random Effects** | | **Variance** | ***SD*** | - | | |
| Participant (intercept) | | 92.72 | 9.63 | - | | |
| Passage (intercept) | | 23.93 | 4.89 | | | |
| Multimodal contrast by Participant (slope) | | 3.30 | 1.82 | | | |
| Integration contrast by Participant (slope) | | 11.08 | 3.33 | | | |
| Presentation language by Participant (slope) | | 7.28 | 2.70 | | | |
| **Random Parameters correlations** | - | | | | | **Correlation** |
| Participant (intercept) & Multimodal contrast | - | | | | | -.113 |
| Participant (intercept) & Integration contrast | | | | | | .221 |
| Participant (intercept) & Presentation language | | | | | | .208 |
| Multimodal contrast & Integration contrast | | | | | | .396 |
| Multimodal contrast & Presentation language | | | | | | -.077 |
| Integration contrast & Presentation language | | | | | | .686 |

*Notes.* Multimodal contrasts showed effects of plain text subtracting the average effects of integrated and separate modes, while the integration contrasts showed effects of separate mode subtracting effects of integrated mode. The presentation language contrasts showed estimate of L1 effects subtracting the average effects of L1 and L2.

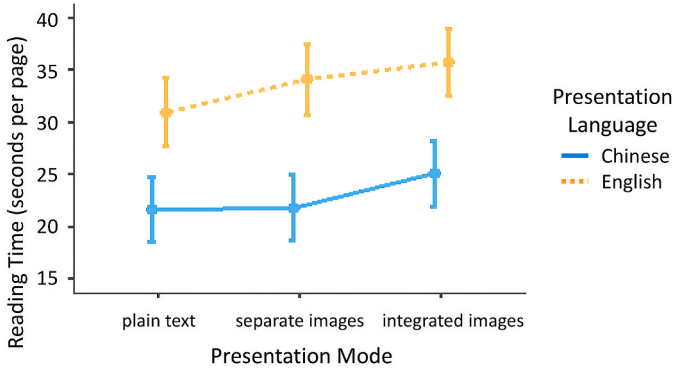


**Fig. 3.** The interaction effects of presentation language and mode on reading time. The error bars indicate 95% Confidence Interval.

easier but there was no corresponding improvement in reading performance. This dissociation of perceived ease and actual reading performance in L2 indicated that participants were inaccurate at monitoring their own level of comprehension, as in the multimedia heuristics (e.g., Lindner et al., 2018; Serra & Dunlosky, 2010). The asymmetry between L1 and L2 ease perceptions was particularly striking because participants rated L2 passages as significantly easier, albeit taking about 50% longer time to read and showing overall similar levels of comprehension performance. Results suggested that readers may be more susceptible to the multimedia heuristics in L2 reading, but this hypothesis warrants further research to confirm.

Direct effects of separate images in L1 text were not significant, indicating that interest and ease fully mediated the multimodal effects on reading time and comprehension. Thus, the presence of separate images caused affective changes in the appraisal of the passage for skilled readers, but not via cognitive operations that linked visual or conceptual representations of text and pictures. In contrast, L2 text with separate images was the condition with the worst reading outcomes overall, with direct effects that led to longer reading time and lower comprehension accuracy relative to plain text. Although the separately presented images were related to the text (not merely decorative) and identical to those in the integrated mode condition, the formatting of the text and image appeared to have caused interference in reading comprehension. Due to higher cognitive efforts involved in reading in L2, readers may seek out images for support in understanding the contents. However, splitting attention between the text and image across the spatial distance and linking the matching text and pictures both involved high cognitive load. So, the independent presentation mode and non-native language may have additive effects that overloaded working memory (Sweller, 2005), causing poor information encoding or retention even though reading time was long.

For visual narratives with integrated images, direct effects were similar across L1 and L2 and demonstrated that the integrated formatting of text and image led to longer reading time independently of situational interest, perceived ease, and other covariates. The bundled nature of integrated text and images likely caused readers to process the images in tandem with the text. Most theories posit that comprehending multimodal messages involves reconciliation of those varying modalities into an integrated understanding (Cohn, 2016; Mayer, 2009; Schnotz, 2005). Regardless of how these mechanisms are ordered, multimodal comprehension will involve cognitive processes such as visual integration, where readers first select the relevant text and image, and conceptual integration, where readers organize and integrate the semantic information in working memory. Such processes will by necessity lead to longer viewing times as information is distributed across additional modalities.

### 4.2. Limitations and future directions

Taken together, the pattern of results supported a unitary system encompassing text and images in reading comprehension. Results clarified that situational interest and perceived ease mediated some effects of multimodal presentation. Direct effects independent from these factors could further facilitate or hinder reading comprehension of narratives, depending the extent of text-picture integration and the

**Table 5**
Parameter estimate results of the generalized mixed-effects model for comprehension accuracy.

| Model equation: Accuracy ~ Mode x Presentation language + Familiarity + Language grade + Visual language fluency + (1 \| Participant) + (1 \| Passage) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Model fit: $R^2$ marginal = .039, $R^2$ conditional = .147, AIC = 1150.04 | | | | | | | |
| Fixed Effects | $\beta$/exp($\beta$)[a] | | *SE* | 95% CI for $\beta$ | | *Z* | *p* |
| Multimodal contrast | 0.18/ | 1.20 | 0.16 | [-0.13 | 0.49] | 1.15 | .249 |
| Integration contrast | −0.35/ | 0.71 | 0.18 | [-0.70 | 0.00] | −1.95 | .052 |
| Presentation language | −0.05/ | 0.95 | 0.15 | [-0.20 | 0.09] | −0.70 | .487 |
| Visual language fluency | −0.18/ | 0.84 | 0.09 | [-0.36 | 0.00] | −1.94 | .053 |
| Language grade | −0.08/ | 0.93 | 0.08 | [-0.24 | 0.09] | −0.89 | .373 |
| Familiarity of the topic | 0.24/ | 1.27 | 0.08 | [0.08 | 0.40] | 2.93 | .003* |
| Multimodal contrast x Presentation language | 0.05/ | 1.05 | 0.16 | [-0.25 | 0.36] | 0.34 | .736 |
| Integration contrast x Presentation language | 0.44/ | 1.55 | 0.18 | [0.09 | 0.79] | 2.44 | .015* |
| **Random Effects** | **Variance** | ***SD*** | | | | | |
| Participant (intercept) | 0.51 | 0.26 | | | | | |
| Passage (intercept) | 0.40 | 0.16 | | | | | |

*Notes.* Multimodal contrasts showed effects of plain text subtracting the average effects of integrated and separate modes, while the integration contrasts showed effects of separate mode subtracting effects of integrated mode. The presentation language contrasts showed estimate of L1 effects subtracting the average effects of L1 and L2.
[a] The exponentiation of the estimates, exp(B), is the odds ratio, provided for easier interpretation in addition to the logged odds ratio.

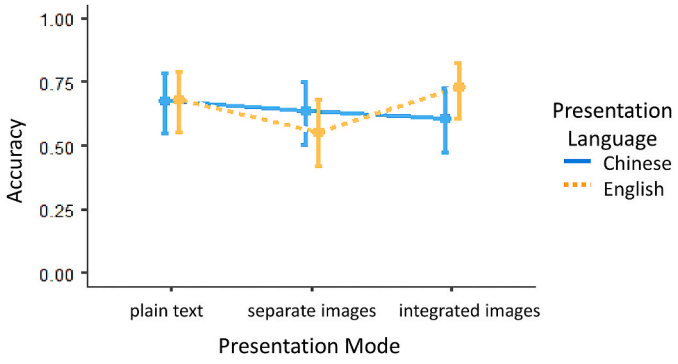


**Fig. 4.** The interaction effects of presentation language and mode on comprehension accuracy. The error bars indicate 95% Confidence Interval.

presentation language. We proposed that the facilitation occurred on the levels of visual and conceptual integration. However, the current study did not include measures that could separately tap into these two processes. Future studies using eye-tracking or neurocognitive methods could more appropriate address the origins of the direct effects.

The rating of perceived ease may be affected by both extrinsic cognitive load induced by instructional design (e.g., multimodal formatting) and intrinsic cognitive load such as prior knowledge or language difficulty that is inherent in the passage contents (Sweller et al., 1998). While the counterbalanced design and statistical covariate of topic familiarity was used to control for difference in intrinsic load in the current study, more specifically worded ratings may differentiate intrinsic and extrinsic cognitive loads. This may further illustrate how reading comprehension is affected by instructional design versus presentation language or familiarity with passage topic. Another limitation of the study was the number of trials used in the experiment. In this study, we only used two multiple choice questions per passage to assess

**Table 6**
Mediation effects of perceived ease and reading interest vs direct effects of presentation modes on reading performance.

| | Chinese reading comprehension | | | English reading comprehension | | |
|---|---|---|---|---|---|---|
| | Estimates | 95% *CI* | *p* | Estimates | 95% *CI* | *p* |
| **Reading time (seconds)** | | | | | | |
| **Separate images**[a] | | | | | | |
| Mediation of Interest | 0.21 | [0.01 0.48] | .037* | 0.25 | [0.01 0.58] | .041* |
| Mediation of Ease | −0.37 | [-0.66 -0.06] | .024* | −0.004 | [-0.14 0.12] | .915 |
| Direct effects | 0.25 | [-1.35 1.86] | .765 | 2.77 | [0.95 4.54] | <001* |
| **Integrated images**[a] | | | | | | |
| Mediation of Interest | −0.12 | [-0.32 0.00] | .071 | −0.10 | [-0.31 0.03] | .170 |
| Mediation of Ease | −0.48 | [-0.90 -0.09] | .015* | −0.18 | [-0.46 0.05] | .110 |
| Direct effects | 3.84 | [2.26 5.42] | <001* | 4.95 | [3.08 6.89] | <001* |
| **Comprehension accuracy (log odds)** | | | | | | |
| **Separate images**[a] | | | | | | |
| Mediation of Interest | 0.004 | [-0.01 0.02] | .450 | −0.001 | [-0.02 0.01] | .729 |
| Mediation of Ease | 0.02 | [0.002 0.04] | .023* | 0.001 | [-0.01 0.01] | .991 |
| Direct effects | −0.08 | [-0.19 0.03] | .136 | −0.11 | [-0.21 -0.01] | .040* |
| **Integrated images**[a] | | | | | | |
| Mediation of Interest | −0.002 | [-0.01 0.001] | .510 | 0.001 | [-0.01 0.01] | .890 |
| Mediation of Ease | 0.03 | [0.003 0.05] | .025* | 0.01 | [-0.001 0.03] | .110 |
| Direct effects | −0.08 | [-0.19 0.03] | .123 | 0.05 | [-0.05 0.14] | .370 |

Notes.
[a] These effects were relative to the plain text condition.

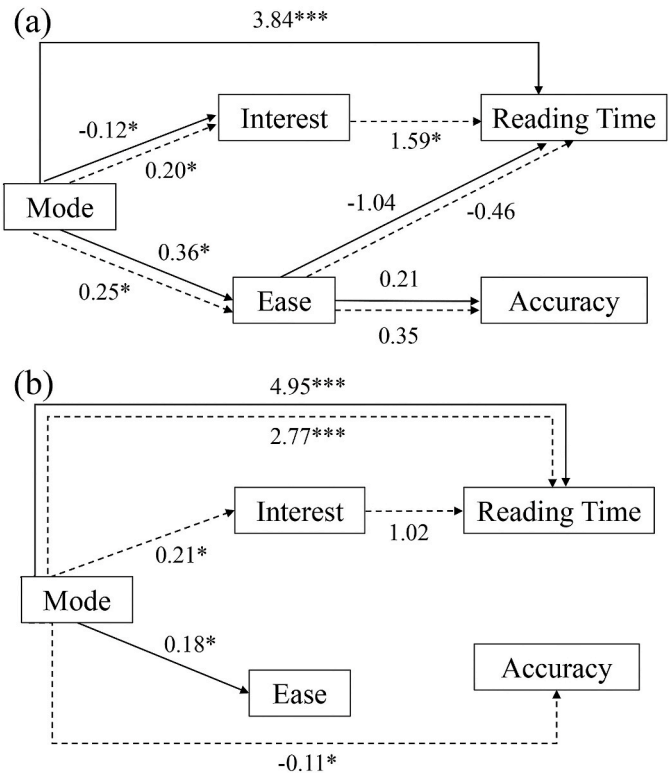


**Fig. 5.** Path diagrams illustrating the direct and indirect causal pathways between presentation modes and reading outcomes in (a) L1 passage reading and (b) L2 passage reading. Solid lines represent pathways in the integrated presentation mode and dashed lines represent pathways in the separate presentation mode.

comprehension accuracy. The Spearman-Brown formula for predicted reliability was applied to show that increasing the number of questions to four per passage would achieve mean reliability of 0.70, and seven questions would be needed for reliability above 0.80. Increasing the number of comprehension questions will give greater reliability and statistical power to detect small effects, for example those in the response times of comprehension questions. Varying the type of questions, such as using short answer questions or questions requiring inferences, may also provide different angles in understanding the comprehension process.

The design of the current study allowed us to comment more broadly on the architecture of language and the communicative system by comparing multimodal effects in the first and second languages of bilingual participants. Current trends in second language acquisition such as Content and Language Integrated Learning create challenges and opportunities for learners to flexibly use language skills, strategies, and multimedia learning materials. The intersection of multimodal reading and bilingual processing would be fertile ground for future research. While comics in education and especially language learning is becoming widely accepted, it should not be viewed as a magic bullet that improve learning performance for everyone. Visual language fluency may influence the degree to which learners may be able to access the information (see Cohn, 2020a), and thus may modulate the benefit of integrated multimodal expressions for a learner. Teachers may thus need to consider developing visual language fluency in their students, guiding learners on the design features and interpretations inherent in text-image integration to maximize their benefits. Many factors moderate the multimedia effect and learners need to be aware of the complexity to get the communicative benefits and avoid the multimedia heuristic. Self-study and leisure reading represent important avenues of student-centered and independent learning. In these cases, affective and metacognitive factors such as situational interest and perceived ease become central to learning motivation and self-monitoring of learning, and should be considered together with presentation format and language.

## Funding

This study was supported by start-up research grant, The Education University of Hong Kong, RG81/2015-2016R.

Disclosure of interest: N.C. has consulted for projects that promote the communicative potential for comics in learning, including language learning. The other authors report no conflict of interest.

## CRediT authorship contribution statement

**Yen Na Yum:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Writing - original draft, Writing - review & editing. **Neil Cohn:** Conceptualization, Methodology, Resources, Visualization, Writing - original draft, Writing - review & editing. **Way Kwok-Wai Lau:** Data curation, Visualization, Writing - review & editing.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2020.101397.

## References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59* (4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-8*. http://CRAN.Rproject.org/package=lme4/.

Brysbaert, M., & New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.1371/10.3758/BRM.41.4.977.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS One, 5*(6), Article e10729. https://doi.org/10.1371/journal.pone.0010729.

Christensen, R. H. B. (2019). *ordinal: Regression models for ordinal data* R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal/.

Coderre, E. L., Cohn, N., Slipher, S. K., Chernenok, M., Ledoux, K., & Gordon, B. (2018). Visual and linguistic narrative comprehension in autism spectrum disorders: Neural evidence for modality-independent impairments. *Brain and Language, 186*, 44–59. https://doi.org/10.1016/j.bandl.2018.09.001.

Cohn, N. (2013). Beyond speech balloons and thought bubbles: The integration of text and image. *Semiotica, 2013*(197), 35–63. https://doi.org/10.1515/sem-2013-0079.

Cohn, N. (2016). A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition, 146*, 304–323. https://doi.org/10.1016/j.cognition.2015.10.007.

Cohn, N. (2020a). Visual narrative comprehension: Universal or not? *Psychonomic Bulletin & Review, 27*(2), 266–285. https://doi.org/10.3758/s13423-019-01670-1.

Cohn, N. (2020b). Your brain on comics: A cognitive model of visual narrative comprehension. *Topics in Cognitive Science, 12*(1), 352–386. https://doi.org/10.1111/tops.12421.

Cohn, N., & Magliano, J. P. (2020). Editors' introduction and review: Visual narrative research: An emerging field in cognitive science. *Topics in Cognitive Science, 12*(1), 197–223. https://doi.org/10.1111/tops.12473.

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea) nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology, 65*(1), 1–38. https://doi.org/10.1016/j.cogpsych.2012.01.003.

Eitel, A., Scheiter, K., & Schüler, A. (2013). How inspecting a picture affects processing of text in multimedia learning. *Applied Cognitive Psychology, 27*(4), 451–461. https://doi.org/10.1002/acp.2922.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532.

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction, 16*(6), 511–525. https://doi.org/10.1016/j.learninstruc.2006.10.001.

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. https://doi.org/10.1006/ceps.1998.0987.

Hong Kong Examinations and Assessment Authority. (2013). *Results of the benchmarking study between IELTS and HKDSE English language examination*. http://www.hkeaa.edu.hk/DocLibrary/MainNews/press_20130430_eng.pdf.

Hong Kong Examinations and Assessment Authority. (2018). *Grading procedures and standards-referenced reporting in the HKDSE*. Retrieved from the Hong Kong Examinations and Assessment Authority website: http://www.hkeaa.edu.hk/DocLibrary/Media/Leaflets/HKDSE_SRR_A4booklet_Mar2018.pdf.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(4), 765–789. https://doi.org/10.1017/S0003055411000414.

International English Language Test System. (2020, June 3). Common European framework: How should the CEFR be used by recognising institutions wishing to set language ability requirements?. https://www.ielts.org/ielts-for-organisations/common-european-framework.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*, 601–625. https://doi.org/10.1146/annurev-psych-122414-033702.

Kirtley, C., Murray, C., Vaughan, P. B., & Tatler, B. W. (2018). Reading words and images: Factors influencing eye movements in comic reading. In *Empirical comics research* (pp. 264–283). Routledge. https://doi.org/10.4324/9781315185354-13.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Laubrock, J., Hohenstein, S., & Kümmerer, M. (2018). Attention to comics: Cognitive processing during the reading of graphic literature. In *Empirical comics research* (pp. 239–263). Routledge. https://doi.org/10.4324/9781315185354-12.

Leutner, D. (2014). Motivation and emotion as mediators in multimedia learning. *Learning and Instruction, 29*, 174–175. https://doi.org/10.1016/j.learninstruc.2013.05.004.

Lindner, M. A., Eitel, A., Barenthien, J., & Köller, O. (2018). An integrative study on learning and testing with multimedia: Effects on students' performance and metacognition. *Learning and Instruction, 101100*. https://doi.org/10.1016/j.learninstruc.2018.01.002.

Liu, J. (2004). Effects of comic strips on L2 learners' reading comprehension. *Tesol Quarterly*, 225–243. https://doi.org/10.2307/3588379.

Magliano, J. P., Larson, A. M., Higgs, K., & Loschky, L. C. (2016). The relative roles of visuospatial and linguistic working memory systems in generating inferences during visual narrative comprehension. *Memory & Cognition, 44*(2), 207–219. https://doi.org/10.3758/s13421-015-0558-7.

Magner, U. I., Schwonke, R., Aleven, V., Popescu, O., & Renkl, A. (2014). Triggering situational interest by decorative illustrations both fosters and hinders learning in computer-based learning environments. *Learning and Instruction, 29*, 141–152. https://doi.org/10.1016/j.learninstruc.2012.07.002.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*(1), 1–19. https://doi.org/10.1207/s15326985ep3201_1.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678.

Merc, A. (2013). The effect of comic strips on EFL reading comprehension. *The International Journal of New Trends in Education and Their Implications, 4*(1), 54–64.

OECD. (2019). *PISA 2018 Reading Framework", in PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing. https://doi.org/10.1787/5c07e4f1-en.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie, 45*(3), 255. https://doi.org/10.1037/h0084295.

Palmer, S. (1992). Common region: A new principle of perceptual grouping. *Cognitive Psychology, 24*, 436–447. https://doi.org/10.1016/0010-0285(92)90014-S.

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review, 1*, 29–55. https://doi.org/10.3758/BF03200760.

Pan, H., Liu, S., Miao, D., & Yuan, Y. (2018). Sample size determination for mediation analysis of longitudinal data. *BMC Medical Research Methodology, 18*(1), 32. https://doi.org/10.1186/s12874-018-0473-2.

Piedmont, R. L. (2014). Inter-item correlations. *Encyclopedia of quality of life and well-being research*. https://doi.org/10.1007/978-94-007-0753-5_1493.

Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Schiefele, U. (2009). Situational and individual interest. In K. R. Wenzel, & A. Wigfield (Eds.), *Educational psychology handbook series. Handbook of motivation at school* (pp. 197–222). Routledge/Taylor & Francis Group.

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). Explaining the modality and contiguity effects: New insights from investigating students' viewing behaviour. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 24*(2), 226–237. https://doi.org/10.1002/acp.1554.

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.005.

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction, 13*, 141–156. https://doi.org/10.1016/S0959-4752(02)00017-8.

Schnotz, W., & Wagner, I. (2018). Construction and elaboration of mental models through strategic conjoint processing of text and pictures. *Journal of Educational Psychology, 110*(6), 850–863. https://doi.org/10.1037/edu0000246.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory, 18*(7), 698–711. https://doi.org/10.1080/09658211.2010.506441.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.003.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296. https://doi.org/10.1023/A:1022193728205.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). *Mediation: R package for causal mediation analysis*. https://doi.org/10.18637/jss.v059.i05.

Wong, S. W., Miao, H., Cheng, R. W. Y., & Yip, M. C. W. (2017). Graphic novel comprehension among learners with differential cognitive styles and reading abilities. *Reading & Writing Quarterly, 33*(5), 412–427. https://doi.org/10.1080/10573569.2016.1216343.

Zhao, F., & Mahrt, N. (2018). Influences of comics expertise and comics types in comics reading. *International Journal of Innovation and Research in Educational Sciences, 5*(2), 218–224.