



Texts and pictures serve different functions in conjoint mental model construction and adaptation

Fang Zhao¹ · Wolfgang Schnotz² · Inga Wagner² · Robert Gaschler^{1,3}

Published online: 1 August 2019
© The Psychonomic Society, Inc. 2019

Abstract

In this study we examined the different functions of text and pictures during text–picture integration in multimedia learning. In Study 1, 144 secondary school students (age = 11 to 14 years; 72 females, 72 males) received six text–picture units under two conditions. In the *delayed-question* condition, students first read the units without a specific question (no-question phase), to stimulate initial coherence-oriented mental model construction. Afterward the question was presented (question-answering phase), to stimulate task-adaptive mental model specification. In the *preposed-question* condition, students received a specific question from the beginning, stimulating both kinds of processing. Analyses of the participants' eye movement patterns confirmed the assumption that students allocated a higher percentage of available resources to text processing during the initial mental model construction than during adaptive model specification. Conversely, students allocated a higher percentage of available resources to picture processing during adaptive mental model specification than during the initial mental model construction. In Study 2 ($N = 12$, age = 12 to 16; seven females, five males), we ruled out that these findings were due to the effect of rereading, by implementing a no-question phase either once or twice. To sum up, texts seem to provide more explicit conceptual guidance in mental model construction than pictures do, whereas pictures support mental model adaptation more than text does, by providing flexible access to specific information for task-oriented updates.

Keywords Text–picture integration · Eye tracking · Initial mental model construction · Adaptive mental model specification

Text accompanied by static pictures is ubiquitous in textbooks, especially those for the natural sciences. Abundant research has shown that students learn better from text *and* pictures than from text alone (e.g., Carney & Levin, 2002; DeLeeuw & Mayer, 2008; Mayer, 2009). Nevertheless, it is not yet well understood how text and pictures interact in their conjoint processing (Ortegren, Serra, & England, 2015).

When both text and picture are needed for comprehension and learning, students must integrate verbal and pictorial information into one coherent, task-appropriate mental

representation, a process known as *text–picture integration*. An example of the need for text–picture integration is presented in Fig. 1, which originates from a textbook on biology. The text describes the dynamic processes of blood circulation between mother and fetus during pregnancy, which is shown in pictures that point out the main parts using numbers. Global understanding, as well as the answering of specific questions, requires students to integrate the text and picture information.

Referring to Wainer's (1992) taxonomy, integration requirements can differ in terms of their complexity. Low-complexity questions require only element mappings between text and picture. For example, to answer the question what is the name of the pink area?, students have to scan the picture; find the pink area, which refers to "3"; search for 3 in the text; and find the correct answer, "placenta." Medium-complexity questions require mappings of simple relations. For instance, students could be asked which parts do not directly link to each other. To find the correct answer, they have to map the blood vessels of the mother and child, the cervix and amniotic fluid, the umbilical cord and amniotic sac, as well as the placenta and uterine wall in the text. Then they have to identify the parts in the picture via number coding, which allows them

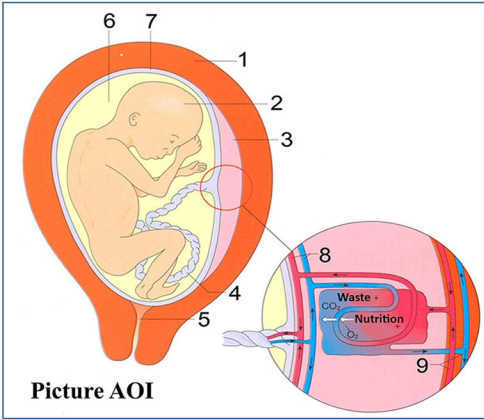
Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13421-019-00962-0>) contains supplementary material, which is available to authorized users.

✉ Fang Zhao
fang.zhao@fernuni-hagen.de

¹ University of Hagen, Hagen, Germany

² University of Koblenz-Landau, Landau, Germany

³ Research Cluster Image Knowledge Gestaltung at Humboldt University of Berlin, Berlin, Germany



Picture AOI

Which parts do not directly link to each other? **Question AOI**

blood vessels of the mother and blood vessels of the child

umbilical cord and amniotic sac

cervix and amniotic fluid

placenta and uterine wall

Text AOI

A child develops inside the uterus, or uterine wall (1). From the fourth month of gestation, it is called a fetus (2). The fetus is nourished by the placenta (3). It is in the placenta that the exchange between the blood vessels of the fetus (8) and the blood vessels of the mother (9) takes place (look at zoomed area). In the blood vessels, nutrients and oxygen (O₂) as well as waste products and carbon dioxide (CO₂) are exchanged. The fetus is connected to the mother by the umbilical cord (4). The amniotic fluid (6) protects the child. You could call it a protective -pillow for the fetus, because it helps to cushion the fetus from impact. When the amniotic sac (7) is broken, the delivery process is initiated and the fetus will move through the cervix (5) in the course of the labour.

Fig. 1 Example of a text–picture integration task on the topic of pregnancy (translated from German)

to answer from the pictures that the cervix and amniotic fluid do not directly link to each other. High-complexity questions require mappings of more complex relations between the text and picture. For instance, which path does the blood of the fetus take after getting nutrition and oxygen (O₂) from the mother? Students cannot answer these questions on the basis of only the text or only the picture. They have to integrate information from the text (which refers to the names and to the number coding) and information from the picture (which provides spatial information and numbers). Then they can answer that blood flows back through the placenta and via the umbilical cord to the fetus.

We assume that conjoint integrative processing of text and static pictures is a goal-directed activity. Goals can vary on a continuum ranging from general to specific. A *general* goal is simply to understand what is explained in the text and shown in the pictures, which means constructing a coherent mental representation for its own sake. A *specific* goal is to solve certain problems or answer certain questions requiring specific information. In the theory of relevance in reading, McCrudden and Schraw (2007) proposed a distinction between text-based importance and task-oriented relevance. *Importance* is the degree to which text segments include information required to understand the text, whereas *relevance* is the degree to which text segments include information that is necessary to perform specific tasks. Following this theory, we assume that students also process texts and pictures differently depending on the task demands. They initially use semantic coherence among texts and pictures (structural importance) as a default orientation for global understanding, but

switch to relevance-oriented processing when they have established specific reading goals to meet particular task demands. Task-oriented processing requires metacognitive monitoring (Vidal-Abarca, Mañá, & Gil, 2010), which includes understanding the task (e.g., a question) and deciding whether to rely on one's prior knowledge or whether a search for information in the text and picture is needed. If a search takes place, cognitive resources have to be allocated appropriately to access, select, process, and integrate the relevant information into the learner's knowledge structure (Britt et al., 2018).

Relevance can be manipulated by instructions, such as to read for a specific purpose or to adopt a specific perspective during reading (Pichert & Anderson, 1977), or by questions that require inferences or elaborative interrogation of the material (Rickards & Denner, 1978; Rouet, 2006), or by priming (van der Laan, Papiés, Hooze, & Smeets, 2017). Rickards (1979) has shown that goal-dependent processing includes increased attention to specific information as well as general backward and forward processes that stimulate mental review processes. Whereas various studies have demonstrated that text processing is highly dependent on the goals of the learner, not much is known about the goal dependency of text–picture integration. In the present article, we aim to analyze whether texts and pictures play different roles during text–picture integration in multimedia learning.

Theory

Texts and pictures implement different forms of representations, which accomplish different functions in the process of

comprehension. Not only can texts and pictures complement each other informationally; texts can also constrain the interpretation of pictures, and pictures can constrain the interpretation of texts (Ainsworth, 1999). As has been demonstrated by Larkin and Simon (1987), texts provide information more explicitly than pictures, whereas pictures provide higher computational efficiency for drawing inferences and problem solving (Gyselinck & Tardieu, 1999). Pictures can especially serve a scaffolding function for constructing mental representations, even after having been presented for just a few seconds, which is not possible with texts (Eitel & Scheiter, 2015; Eitel, Scheiter, Schüler, Nyström, & Holmqvist, 2013; Lindner, Eitel, Strobel, & Köller, 2017). Despite the various beneficial effects of pictures, eye movement studies have shown that multimedia reading is heavily driven by the texts and only minimally by the pictures (e.g., Hannus & Hyönä, 1999, Exp. 2).

Theoretical approaches to multimedia learning

Dual-coding theory (DCT) A common view of multimedia learning is the one represented by the dual-coding theory of Paivio (1986), referred to by Kulhavy, Lee, and Caterino (1985) in their conjoint-processing theory. The basic assumption is that text and pictures are processed in two cognitive sub-systems: A verbal and a pictorial system. Verbal information is processed and encoded only in the verbal system, whereas pictorial information is processed and encoded both in the pictorial and in the verbal system.

Parallel multimodal architecture (PMA) Cohn (2016) has recently proposed a theoretical framework assuming that our cognitive system perceives the information through different modalities (e.g., verbal, visual) and different decoding structures (e.g., text for syntax, sequential images for narrative), and decides whether one type of information is dominant.

Cognitive theory of multimedia learning (CTML) The CTML proposed by Mayer (2009, 2014), which is partially inspired by the dual-coding theory, provides a more differentiated view. CTML states that a working memory of limited capacity has an auditory–verbal channel for processing texts and a visual–pictorial channel for processing pictures. Processes of selection and organization result in a verbal mental model (or knowledge structure) within the auditory–verbal channel, and in a pictorial mental model within the visual–pictorial channel. The two mental models are then integrated into a coherent and more elaborate mental representation. Because texts and pictures complement one another, comprehension is enhanced when learners are able to mentally integrate verbal and pictorial information.

Integrative model of text–picture comprehension (ITPC model) Another theoretical approach to multimedia learning,

which puts more emphasis on the representational differences between texts and pictures, is represented by the ITPC model (Schnotz, 2014; Schnotz & Bannert, 2003). Integrative processing here is assumed to take place in a verbal (i.e., descriptive) channel and in a pictorial (i.e., depictive) channel. The verbal channel includes the external text, the internal text surface representation, and the propositional representation of the semantic content of the text. Information processing in this channel occurs by means of symbol processing. The pictorial channel involves the external picture, the internal visual image of the picture, and the mental model of the subject matter. In accordance with Johnson-Laird (1983), a mental model is viewed as any mental representation based on analogy. Information processing in the pictorial channel therefore takes place by structure mapping based on analogies (i.e., structural correspondences) between the depictive representations (Gentner, 1989; Knauff & Johnson-Laird, 2002; Sims & Hegarty, 1997). The ITPC model assumes continuous interactions between propositional representations and mental models in terms of mental model construction and mental model inspection processes.

Processing constraints and information access Because the ITPC model puts special emphasis on the representational differences between texts and pictures, it allows for making predictions about functional differences between both kinds of representations in multimedia learning. For this reason, we used the ITPC model as the theoretical framework for our study. Expository texts are considered in the ITPC model to be descriptive representations with an inherent linear structure. Students are expected to read texts word by word, in a predetermined order. The processing of text is thus highly constrained in terms of processing order. Because expository texts explicitly deliver specific propositions in a particular sequence, their informational content is relatively well defined. This makes texts especially well adapted to provide conceptual guidance for the process of comprehension. Single static pictures in expository text do not possess such an inherent linear structure.¹ Although a picture can include salient visual elements or can convey information about sequences, there is usually no predetermined order of information processing within the picture (Massironi, 2002; Sadoski & Paivio, 2001). Although pictures allow for generating propositions by reading off corresponding information, their propositional content is not clearly defined. It is only implicit, depending on the procedures applied to the picture. Accordingly, the processing of pictures is less constrained in terms of processing order. Their semantic content is less

¹ We do not deny that linear structures can play a role in picture comprehension—when, for example, visual narratives such as comics present sequences of pictures that require an analysis of relations between subsequent pictures. However, our focus here is on single static, explanatory pictures.

clearly defined than the semantic content of texts, which makes them poorer conceptual guides than texts.

When task-oriented processing requires searches for specific information, texts and pictures differ in their information accessibility. The linear structure of texts makes searches for specific information more difficult, insofar as the path to be searched for a specific piece of information is on average longer than with a two-dimensional arrangement of information. Readers can speed up the search by scanning the text, but this includes the risk of overlooking relevant information. Since single static pictures do not have a linear, but a two-dimensional, structure, they can be searched, *ceteris paribus* (all else being equal), more easily for specific information, because the search paths within a picture are shorter than those within a linear structure. All in all, in terms of information access, the structure of text seems to be less suitable for searching for specific information than that of pictures, whereas pictures provide relatively easy and more flexible information access (Schnotz & Wagner, 2018).

Initial mental model construction and adaptive mental model specification

For goal-oriented processing of texts and pictures, we assume, in line with McCrudden and Schraw (2007), that learners initially use a semantic coherence strategy as a default orientation for global understanding, and then switch to relevance-oriented processing after having established specific reading goals to meet particular task demands. Accordingly, if learners study text with pictures without a specific task in mind, they engage in general coherence-oriented processing, resulting in a global understanding of the subject matter that can be elaborated further, if required. Thus, general coherence-oriented processing leads to a globally coherent mental model that is not specialized in tasks. We call this initial task-independent, general coherence-oriented processing *initial mental model construction* (IMC).

On the contrary, if learners study text with pictures in order to complete a specific task, they engage at some point in more selective processing in order to adapt their mental model accordingly, placing special emphasis on task-relevant information. This will lead to a mental model that also includes the specific details required to solve the tasks. We call this kind of processing *adaptive mental model specification* (AMS).

We assume inherent sequential constraints between IMC and AMS: Before learners begin answering a question, they want to have at least some minimal knowledge about the subject matter. Thus, some IMC is required before AMS can occur. Accordingly, AMS is assumed to take place on top of IMC, because a mental model must exist before it can be adapted to specific purposes.

Hypotheses

If conjoint processing of text and pictures requires integration of both kinds of information, for overall understanding as well as for accomplishing certain tasks, this does not mean that text and picture have to carry the same weight in the overall meaning (cf. Cohn, 2016). How much and which information is presented by the text versus by the picture depends on the content, the learner's prior knowledge, and the task requirements. In some cases, the text is semantically more dominant than the picture, whereas in other cases, the picture is semantically more dominant than the text. However, despite these idiosyncrasies, one can expect specific differences between IMC and AMS with regard to the use of text and the use of pictures.

For IMC, a short glance at a picture can be a useful scaffold (Eitel & Scheiter, 2015; Eitel et al., 2013). Beyond that, however, pictures provide little conceptual guidance for systematic mental model construction, due to their less clearly defined semantic content and the absence of a predetermined order of processing. Texts, on the contrary, are more suitable to guide the learner's conceptual analysis of the subject matter, because a text's semantic content is more clearly defined and the order in which it is processed is highly constrained. Therefore, texts provide more explicit guidance for the (task-independent) systematic IMC than do pictures. This leads to the first assumption regarding the different roles of text and pictures during text–picture integration in multimedia learning:

Learners allocate, *ceteris paribus*, a higher percentage of available resources to text processing during initial mental model construction than during adaptive model specification.

As for AMS, learners need to have quick and flexible access to task-relevant information in order to adapt a mental model to specific requirements. Texts seem not to be an asset for searching specific information, due to their linear structure, associated with longer search paths. Pictures, on the contrary, provide faster and more flexible access to specific information. Accordingly, one can expect that

Learners allocate, *ceteris paribus*, a higher percentage of available resources to picture processing during adaptive model specification than during initial mental model construction.

Study 1

In Study 1, we tested the different roles of text and pictures during text–picture integration in multimedia learning. The hypotheses were tested with secondary school students by

analyzing their eye movements during the integrative processing of texts and pictures.

Method

Participants On the basis of an a-priori power analysis, 144 German students were recruited for the experiment. Of these, 72 students were from a lower-grade level (fifth-graders: $M_{age} = 11.4$ years, $SD = 0.6$; 28 females, 44 males). Another 72 students were recruited from higher-grade levels (seventh- and eighth-graders: $M_{age} = 14.5$ years, $SD = 0.7$; 44 females, 28 males). Half of the lower-grade and of the higher-grade students attended a “Gymnasium,” which is an academic-track (AT) school. The other half attended a “Realschule,” which is a non-academic-track (non-AT) school. AT school students are on average more competent in reading and more willing to invest effort into working on cognitive tasks than are non-AT school students (Roeschl-Heils, Schneider, & van Kraayenoord, 2003). Each participant was awarded €12 for participating in Study 1 (40 min).

Experimental material We selected six aligned text–picture combinations (henceforth called “text–picture units”) randomly from authentic geography and biology textbooks for grades 5 to 8 in Germany as the experimental material. The text–picture units and corresponding questions are listed in the Appendix B. The units dealt with the structure of insect legs (67 words), the banana trade (91 words), the auditory ranges of animals and humans (130 words), pregnancy (143 words), the map of Europe (136 words), and the types of savannahs (168 words). They referred to both static spatial arrangements

and dynamic processes. Thus, the text–picture units could be considered ecologically valid and educationally balanced. To test for the generalizability of our hypotheses, we varied the complexity of the questions. All questions could only be answered correctly by combining information from the text and from the picture. As is shown in Fig. 1, all displays included an area of interest (AOI) on the right-hand side including the text, and an AOI on the upper left-hand side including the picture. A third AOI, on the lower left-hand side, remained empty or contained the corresponding question, depending on the experimental treatment.

Experimental treatments To create different processing conditions as manipulation of the independent variable, we implemented a *delayed-question condition* including two phases (see Fig. 2, left panel). In the first phase, participants received a text–picture unit without any question (*no-question phase*). This was meant to stimulate IMC only, because participants had no specific task in mind. In the second phase, they were presented the same text–picture unit just seen before, but now together with a (delayed) question in the corresponding AOI (*question-answering phase*). This condition was meant to stimulate AMS, as participants were expected to engage in task-oriented processing adapting their mental model to the requirements of the specific task.

Furthermore, we implemented a *preposed-question condition* (see Fig. 2, right panel), in which a specific question was presented first, followed by the corresponding text–picture unit, where the question remained visible.

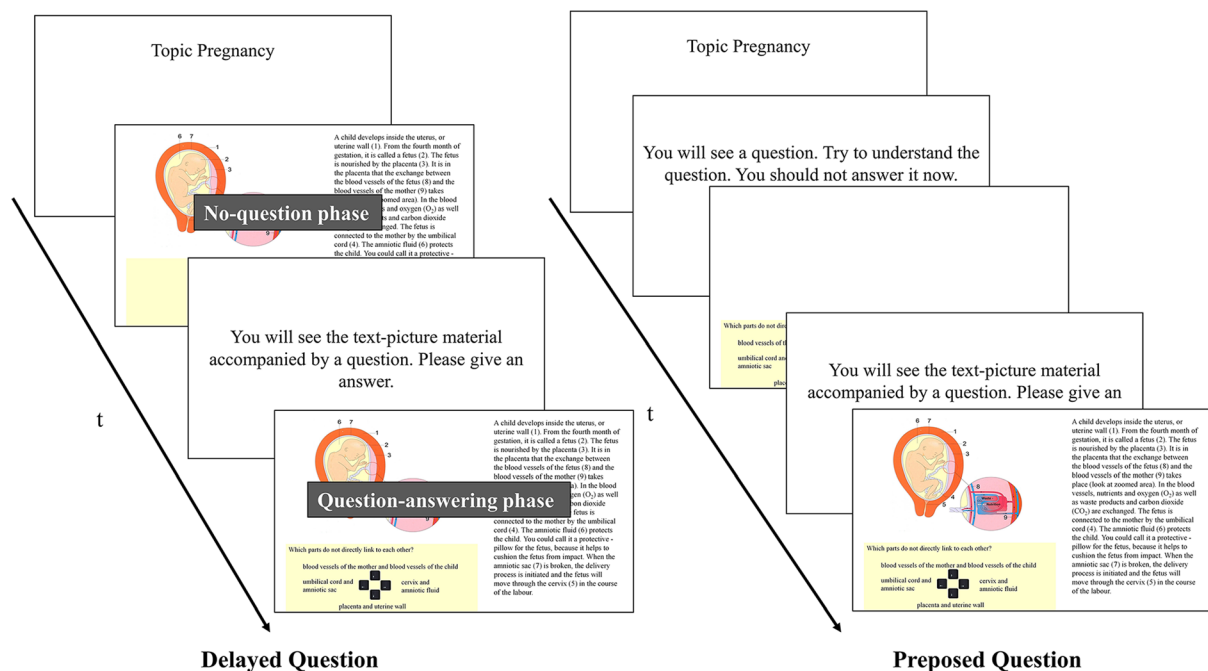


Fig. 2 Display of the reading conditions and phases on the screen for Study 1

participants were free to combine IMC and AMS ad libitum with their more predetermined use of texts and pictures during the no-question and question-answering phases of the delayed-question condition. The experimental conditions were implemented through a within-subjects Latin square design, which counterbalanced the order of experimental conditions (see [Appendix A](#)). For all conditions and for each participant, a text–picture unit was always combined with only one out of the three questions of different complexity.

Procedure and scoring The experiment was conducted individually in a lab environment. First, participants' verbal and spatial intelligence were tested using the German version of the Cognitive Abilities Test (Heller & Perleth, 2000). Then, after a short explanation of how to handle the experimental environment, the learning material was presented with an eye tracker. Presentation of the material was self-paced under all conditions: Students pressed the space key to move from one unit or one condition to the next one, and pressed the up/down/left/right arrows to answer questions. The number of questions answered correctly was recorded for each student.

Participants received six text–picture units and answered six questions. Three units were presented with delayed questions: When the participants had finished their first reading of the text–picture unit in the no-question phase, they switched to the question-answering phase by pressing the space key. The same text–picture unit was displayed as before, but now combined with a specific question. The other three units were presented under the proposed-question condition: When participants had read the question, they pressed the space key that led them to the presentation of the corresponding text–picture unit together with the question.

Eye movements were registered with a Tobii XL60 24-in. eye tracker at 60 Hz. Following a successful calibration, students read text–picture units and answered questions at their own pace. Upon answering the question, the next text–picture unit appeared immediately. They could not return to the previous page. For each participant and each AOI under each condition and phase, the percentages of the accumulated fixation time on text and picture were averaged across the text–picture units (such that the % time on picture and % time on text summed up to 100%). Likewise, the transitions between AOIs were averaged across units. For additional time course analyses, we divided processing phases per participant and unit into quintiles. The total time was set equal to 100%. Within each of the five time intervals, the fixation times for the text, the picture, and the question were determined and expressed as a percentage of the total fixation time. This allowed for analyzing how the participants' visual attention was distributed to the various sources of information during the course of processing, regardless of whether the participants processed the materials slowly or quickly.

Predictions Our assumption about functional differences between the texts and pictures during conjoint processing was tested by analyzing the distribution of fixations under different processing conditions. Visual attention manifests itself in eye fixations, which mirror the bottom-up components of cognitive processing. These bottom-up components are especially dominant during the initial reading of expository material (Just & Carpenter, 1980).

Fixation times depend not only on the amount or intensity of processing, but also on learner expertise. However, our Latin square within-subjects design allowed us to control for learner characteristics, because each participant performed text–picture integration under all three processing conditions, and each text–picture unit was also presented under each condition.

According to our hypothesis, the percentage of text processing versus picture processing would be shifted during IMC toward the text, whereas it would be shifted during AMS toward the picture. We thus predicted for the delayed-question condition that the percentage of accumulated fixation times on the text would be higher in the no-question phase than in the question-answering phase, whereas the percentage of accumulated fixation times on the picture would be higher in the question-answering phase than in the no-question phase (**Prediction 1**).

As for the proposed-question condition, here students' processing could be task-oriented from the beginning. Nevertheless, processing under this condition was expected to include some IMC before AMS takes place, because even highly task-oriented readers need some understanding of what the text–picture unit is about. Thus, we predicted that the percentage of accumulated fixation times on the text would be higher at the beginning of processing and then decrease, whereas the percentage of accumulated fixation times on the picture would be lower at the beginning and then increase (**Prediction 2**).

Because fixations on items indicate AMS, which is assumed, *ceteris paribus*, to be based more on picture than on text processing, we also predicted that there would be more eye movement transitions between picture and question than between text and question during the question-answering phase of the delayed-question condition (**Prediction 3**) and during the proposed-question condition (**Prediction 4**).

Results

Not surprisingly, the AT school students answered more questions correctly than did the non-AT students (61.7% [$SD = 22.3\%$] vs. 49.2% [$SD = 20.8\%$]), $F(1, 140) = 15.36$, $p < .001$, $\eta_p^2 = .10$, and also showed higher verbal intelligence, $F(1, 140) = 26.08$, $p < .001$, $\eta_p^2 = .16$, as well as higher spatial intelligence, $F(1, 140) = 11.11$, $p = .001$, $\eta_p^2 = .07$. Higher-grade students answered more questions correctly than did

Table 1 Percentages of accumulated fixation times on text and pictures in the no-question phase of the delayed-question condition (initial mental model construction)

Variable	Lower-Grade Level		Higher-Grade Level		Total	
	<i>n</i>	<i>M (SD)</i>	<i>n</i>	<i>M (SD)</i>	<i>n</i>	<i>M (SD)</i>
Accumulated Fixation Times: Text/ Picture (%)						
AT School	36	83.3%/16.7% (8.0%)	36	76.5%/23.5% (10.3%)	72	79.9%/ 20.1% (9.8%)
Non-AT School	36	84.8%/15.2% (10.4%)	36	79.9%/20.1% (9.3%)	72	82.4%/17.6% (10.1%)
Total	72	84.1%/15.9% (9.2%)	72	78.2%/21.8% (9.9%)	144	81.1%/18.9% (10%)

For better readability, we report both percentages (adding up to 100%). The bold cell represents the averaged values across all the participants.

lower-grade students (66.1% [*SD* = 19.1%] vs. 44.8% [*SD* = 20.4%]), $F(1, 140) = 45.00, p < .001, \eta_p^2 = .25$. No difference was found between the delayed-question condition and the preposed-question condition, 56.9% (*SD* = 31.2%) vs. 53.9% (*SD* = 28.2%), $p > .35$.

Under the delayed-question condition, the average accumulated fixation times were 48.3 s on the text and 10.6 s on the picture during the no-question phase. The average time on the text during the question-answering phase was 8.4 s, and the average time on the picture was 15.0 s. Students who performed well in answering the questions invested less fixation time on the text during the no-question phase ($r = -.23, p < .01$), but more time on the picture during the question-answering phase ($r = .29, p < .01$). Under the preposed-question condition, the average time on the text was 43.6 s, whereas the average time on the picture was 19.4 s. No significant correlations were found between performance in question-answering and the fixation times on text or pictures. According to our hypotheses, the following analyses will focus on text usage and picture usage under the different processing conditions.² Further data are available at <https://osf.io/qhkpe/>.

Delayed-question condition Table 1 lists the average percentages of accumulated fixation times on texts and pictures for students from different school types and grade levels during the no-question phase. Table 2 shows the corresponding data plus the eye movement transitions between AOIs during the question-answering phase.

A (2×)2×2 analysis of variance (ANOVA) on the percentages of accumulated fixation times on the picture (of the accumulated fixation times on text plus picture), with condition

(no-question phase/ question-answering phase) as a within-subjects factor and school type (AT/non-AT) and grade level (higher/ lower) as between-subjects factors, revealed a significant effect of condition, $F(1, 140) = 779.30, p < .001, \eta_p^2 = .85$. The percentage of text fixation time was higher during the no-question phase than during the question-answering phase, whereas the percentage of picture fixation time was higher during the question-answering phase than during the no-question phase, which confirmed **Prediction 1**. A significant Condition × Grade interaction, $F(81, 140) = 9.61, p = .002, \eta_p^2 = .06$, revealed that lower-grade-level students showed larger differences in picture and text fixation times between the two phases than did higher-grade-level students. No other moderation effects were found.

A (2×)2×2 ANOVA of the eye movement transitions between text, picture, and question during the question-answering phase, with transition type (text–question/picture–question) as a within-subjects factor and school type (AT/non-AT) and grade level (higher/lower) as between-subjects factors, revealed a significant effect of transition type, $F(1, 140) = 167.18, p < .001, \eta_p^2 = .54$. Students showed a higher number of transitions between pictures and questions than between texts and questions, which confirmed **Prediction 3**. The number of correctly answered items was positively related to the number of transitions between text and picture ($r = .32, p < .01$). No moderation effects were found.

Preposed-question condition Since the preposed-question condition involved IMC as well as AMS, it allowed us to estimate the relative contributions of the text and the picture to the overall meaning construction, with the help of accumulated fixation times on the texts and pictures. As is shown in Table 3, the average text contribution is about 67%, and the picture contribution is 33%, and thus in-between the values reported above for the no-question phase (81% vs. 19%) and the question-answering phase (37% vs. 63%) of the delayed-question condition. A 2×2 ANOVA on the percentages of

² The analysis of variance of the percentages of accumulated fixation times on text versus pictures, with complexity of questions and condition as within-subjects factors and school type and grade level as between-subjects factors revealed no interaction of complexity with any other factor. Thus, complexity of the questions was excluded from further analyses.

Table 2 Percentages of accumulated fixation times on text and picture in the question-answering phase of the delayed-question condition (adaptive mental model specification)

Variable	Lower-Grade Level		Higher-Grade Level		Total	
	<i>n</i>	<i>M (SD)</i>	<i>N</i>	<i>M (SD)</i>	<i>n</i>	<i>M (SD)</i>
Accumulated Fixation Times: Text/ Picture (%)						
AT School	36	37.4%/62.6% (16.7%)	36	35.7%/64.3% (13.1%)	72	36.6%/63.4% (14.9%)
Non-AT School	36	33.6%/66.4% (16.5%)	36	42.9%/57.1% (14.4%)	72	38.1%/61.8% (16.1%)
Total	72	35.5%/64.5% (16.6%)	72	39.3%/60.7% (14.1%)	144	37.4%/62.6% (15.5%)
Transition Counts						
Total	72		72		144	
Text–Picture		11.87 (10.06)		17.32 (8.90)		14.62 (9.85)
Picture Question		24.46 (15.48)		26.94 (12.62)		25.70 (14.13)
Text Question		10.65 (7.99)		12.68 (6.85)		11.67 (7.48)

Bold cells represent the averaged values across all the participants.

accumulated picture fixation time under the preposed-question condition, with school type (AT/non-AT) and grade level (higher/lower) as between-subjects factors, showed no main effects of school and grade and no interaction effect, $F_s < 1.14$.

A $(2 \times 2) \times 2$ ANOVA of the eye movement transitions between text, picture, and question under the preposed-question condition, with transition type (text–question/picture–question) as a within-subjects factor and school type (AT/non-AT) and grade level (higher/lower) as between-subjects factors, revealed a significant effect of transition type, $F(1, 140) = 123.73$, $p < .001$, $\eta_p^2 = .47$. Students showed a higher number of transitions between pictures and questions than between

texts and questions, which confirms **Prediction 4**. The number of correctly answered questions was significantly related to the number of transitions between text and picture ($r = .33$, $p < .01$). No other effects were significant.

Time course analyses To test the prediction that the percentage of accumulated fixation times on the text would decrease and that the percentage on the picture would increase under the preposed-question condition, we divided the processing phases per participant and unit into quintiles, introducing the factor of time interval. Figure 3 shows the distributions of fixation times on text, picture, and question over the course of processing under

Table 3 Percentages of accumulated fixation times on text and picture under the preposed-question condition (involving both initial mental model construction and adaptive mental model specification)

Variable	Lower-Grade Level		Higher-Grade Level		Total	
	<i>N</i>	<i>M (SD)</i>	<i>n</i>	<i>M (SD)</i>	<i>n</i>	<i>M (SD)</i>
Accumulated Fixation Times: Text/Picture (%)						
AT School	36	68.6%/31.4% (12.9%)	36	63.6%/36.4% (7.5%)	72	66.1%/33.9% (10.8%)
Non-AT School	36	69.0%/31.0% (20.5%)	36	68.8%/31.2% (10.2%)	72	68.9%/31.1% (16.1%)
Total	72	68.8%/31.2% (17%)	72	66.2%/33.8% (9.3%)	144	67.5%/32.5% (13.7%)
Transition Counts						
Total	72		72		144	
Text–Picture		26.39 (17.35)		39.18 (23.74)		32.78 (21.69)
Picture Question		23.71 (13.86)		25.21 (13.74)		24.46 (13.74)
Text Question		12.35 (9.05)		13.33 (8.03)		12.84 (8.54)

Bold cells represent the averaged values across all the participants.

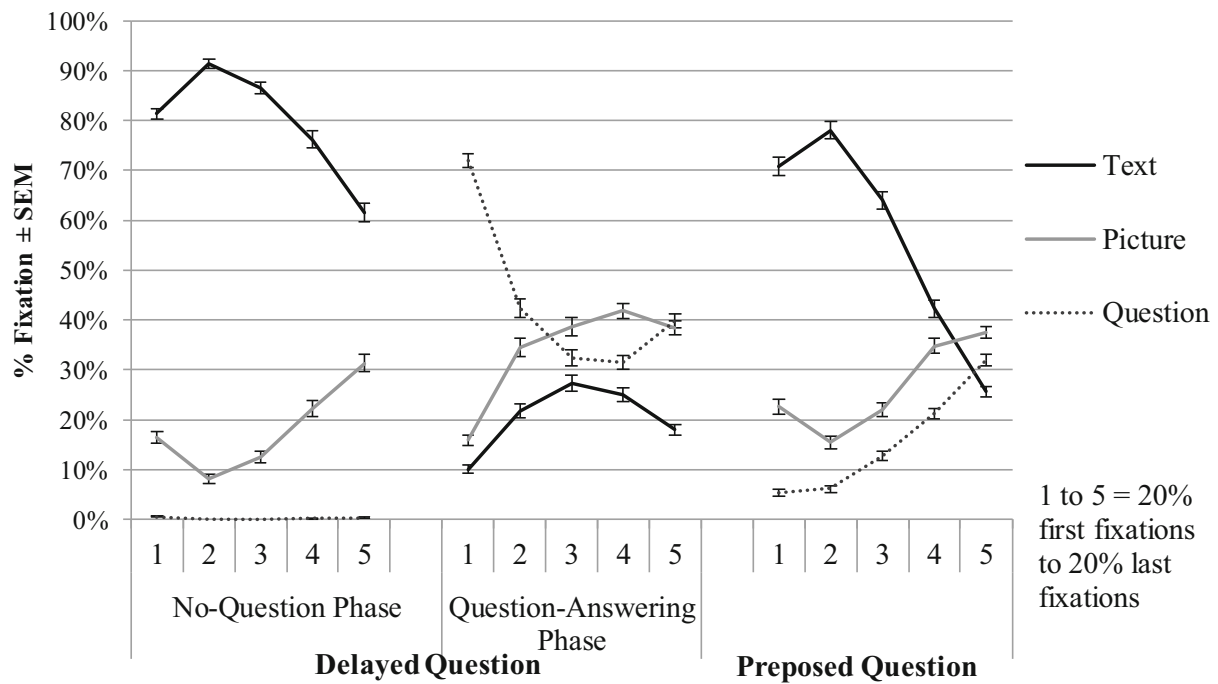


Fig. 3 Percentages of fixation time on the text, picture, and question within five quintiles during the no-question phase and the question-answering phase under the delayed-question condition, and during the preposed-question condition. The error bars indicate the standard errors of the means

different processing conditions, averaged across participants, school types, and grade levels. Since text was dominant during the no-question phase under the delayed-question condition, the data suggest that IMC was more text-driven than picture-driven. However, the relative dominance of text processing gradually decreased during the course of processing, whereas the amount of picture processing gradually increased, probably indicating the termination of IMC. Furthermore, at the very beginning of processing, in the first quintile, the amount of picture processing was a bit higher than in the second quintile. This difference might result from a very short scaffolding function of the pictures at the beginning of IMC. Our time course analysis of the question-answering phase illustrated guidance of processing by the question, especially at the beginning of processing. Contrary to the no-question phase, we observed a higher amount of picture processing than text processing for AMS.

The curves for text processing and for picture processing under the preposed-question condition were found to be similar to the curves of the no-question phase in the first four quintiles, and to the curves of the question-answering phase in the fifth quintile. For the preposed-question condition, Fig. 3 shows an initial increase in the percentage of fixations on the text, with a reduction from the second quintile onward. Conversely, the percentage of fixations on the picture showed an initial reduction, followed by an increase from Quintile 2 onward, $F(3.12, 445.54) = 137.95, p < .001, \eta_p^2 = .49$.³

³ This result is from a one-way ANOVA with the percentage of accumulated fixation time on pictures (of the total fixation time on text and on pictures) as the dependent variable and with the five quintiles as independent variable.

The dominance of text processing supports the assumption that even with preposed questions, processing begins with IMC, which is (*ceteris paribus*) by tendency more text-driven. As is indicated by the increasing percentage of item fixations, AMS takes place on top of the preceding IMC. This confirms our **Prediction 2**, of inherent sequential constraints between IMC and AMS (see Fig. 3).

Discussion

The results of Study 1 confirmed all our predictions. Thus, one could argue that the findings affirm the hypothesis that texts and pictures accomplish different functions in their conjoint processing. Processing is shifted toward the text during IMC, whereas during AMS processing is shifted toward the picture. However, one could also consider the following alternative explanations.

Learners frequently need more than one reading to comprehend difficult texts (Millis & King, 2001). Higher order processes thus cannot be executed before their inputs from subordinate processes are available. Subordinated processes such as word access and the assembly of a propositional representation (*i.e.*, text base) have the highest priority during first reading, because higher order processes depend on their outputs. Rereading then allows learners to complete processes, which were left incomplete during first reading, focusing on higher order processing such as the construction of a mental model. In short: First reading and rereading could allocate cognitive resources to different kinds of processing (Millis, Simon, & tenBroek, 1998).

The texts used in this study were selected from corresponding schoolbooks and did not seem to be especially difficult in terms of length, syntactic or semantic complexity. Furthermore, as participants could read the material self-paced, they had under all experimental conditions and in each phase sufficient time to read and reread the material in order to elaborate, repair, complete, and verify their comprehension, including the construction of a mental model. Nevertheless, AMS during the question-answering phase of the delayed question condition admittedly required another rereading of the material. Thus, one could argue that our findings were not due to the different functions of text and picture during IMC and AMS, but resulted simply from rereading the material. To rule out this alternative explanation, we conducted a second experiment in which a distinction was made between first reading without a question, rereading without a question, and rereading after presenting a question.

Study 2

In Study 2 we examined whether the findings in Study 1 were due to the effect of rereading. Unlike in Study 1, the no-question phase was exposed either once or twice to the participants (see Fig. 4).

Method

Thirteen seventh-graders participated in Study 2, from whom 12 datasets ($M_{\text{age}} = 13$ years, $SD = 1.2$; seven females, five

males) are reported, due to reading difficulties of one participant. Each participant was rewarded €15 for participating in Study 2 (60 min). The experimental materials were the same six text–picture units and associated questions used in Study 1. Three of the text–picture units were presented under a single delayed-question condition, which consisted of a single no-question phase followed by a question-answering phase, just as in the delayed-question condition in Study 1 (see Fig. 4, left panel). The other three text–picture units were presented under a twofold delayed-question condition, consisting of three phases (see Fig. 4, right panel). In the first phase, participants received a text–picture unit without any question (*first no-question phase*). When they had finished reading, they were instructed via the screen to read the text–picture unit again without any question (*second no-question phase*). After they had finished their rereading in this phase, they were presented the same text–picture unit again, but now together with a question they had to answer (*question-answering phase*). The further experimental procedure, including scoring, was performed exactly as in Study 1. An SMI RED 250-Hz system was used for eye tracking. The single versus twofold delayed-question conditions were counterbalanced.

According to the hypotheses in Study 1, the percentage of text processing versus picture processing should be shifted toward the text during IMC, whereas the percentage of text processing versus picture processing should be shifted toward the picture during AMS. We therefore predicted in Study 2 that the percentage of accumulated fixation times on the text would be higher in the no-question phase(s) than in the question-answering phase, whereas the percentage of

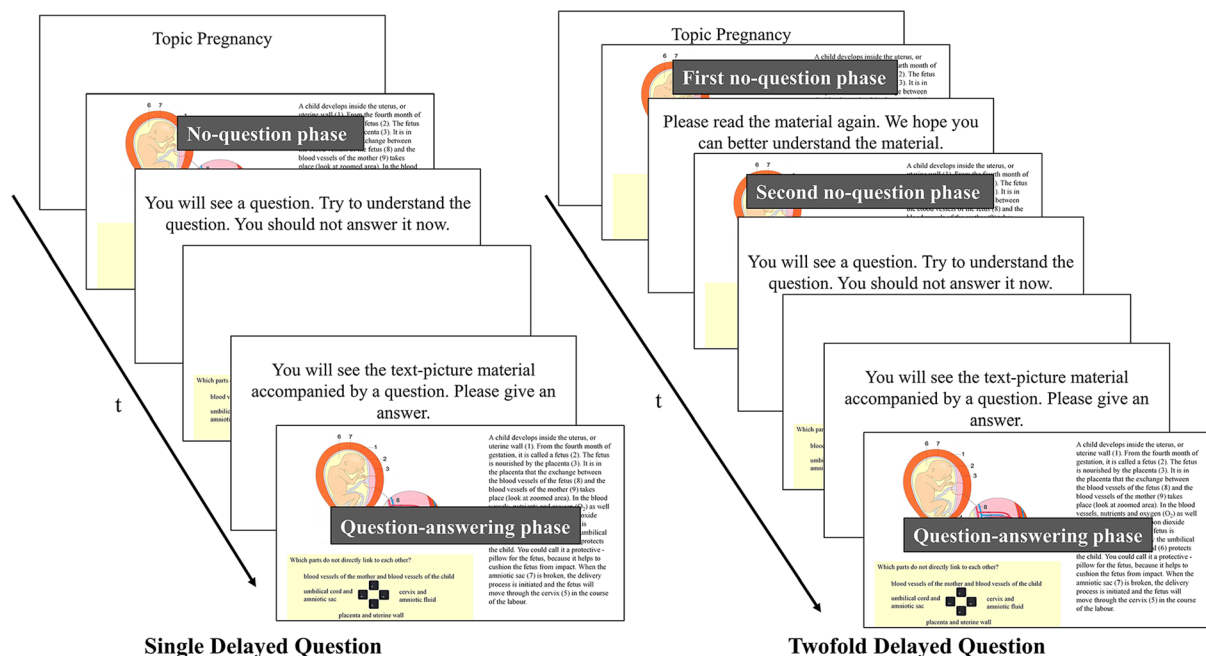


Fig. 4 Display of the reading conditions and phases on the screen for Study 2

accumulated fixation times on the picture would be higher in the question-answering phase than in the no-question phase(s) (**Prediction 5**). We also assumed that the differences between text processing and picture processing were due to the different functions of the text and pictures in IMC and AMS, rather than resulting from rereading the material. We thus further predicted that the difference between the percentage of text or picture processing between the first and second no-question phases would only be small, whereas the difference between the second no-question phase and the following question-answering phase would be much greater (**Prediction 6**).

Results

Participants answered 69.4% (*SD* = 12.2%) correctly in the single delayed-question condition, which was similar to their performance in the twofold delayed-question condition (*M* = 75%, *SD* = 14.4%), *p* = .59. The average correct rate for all reading materials was 71.7% (*SD* = 21.7%). As predicted, we found under the single delayed-question condition a significant decrease of the percentage of text fixations (from 76.6% to 47.9%), and a corresponding increase of the percentage of picture fixations (from 23.4% to 52.1%), *t*(11) = 7.79, *p* < .001, *d* = 2.25, between the no-question phase and the question-answering phase (see Table 4). Under the twofold delayed-question condition, we found no decrease of the percentage of text fixations between the first and second no-question phases. Numerically, the percentage even increased, from 72.4% to 77.6%, and the percentage of picture processing decreased correspondingly from 27.6% to 22.4%, *t*(11) = - 1.49, *p* = .16, *d* = 0.43. However, there was a significant decrease of the percentage of text fixations (from 77.6% to 47.1%) between the second no-question phase and the following question-answering phase, and a corresponding increase of the percentage of picture fixations (from 22.4% to 52.9%), *t*(11) = 6.13, *p* < .001, *d* = 1.77.

Thus, the percentage of picture fixations was higher in the question-answering phase than in the no-question phases (which confirmed **Prediction 5**), and the difference between the second no-question phase and the following question-answering phase was much higher than the difference between the first and second no-question phases (which confirmed **Prediction 6**).

Discussion

All predictions for Study 2 were confirmed. Accordingly, the differences between texts and pictures with regard to processing during a preceding no-question phase and a following question-answering phase seem to have been due to the difference between IMC and AMS rather than to rereading the material. The decrease of the percentage of fixations on text and the increase of the percentage of fixations on pictures from a previous no-question phase to a following question-answering phase under the single and twofold delayed-question condition were nearly the same. Thus, the rereading during the second no-question phase did not seem to affect the results.

General discussion

In the present studies we aimed to clarify the different functions of text and pictures during text–picture integration in multimedia learning. On the basis of the integrated model of text and picture comprehension as a theoretical framework, text and pictures were assumed to play different roles in multimedia learning.

All predictions derived from our hypotheses were confirmed. It seems that initial mental model construction is indeed more (but not exclusively) text-driven than picture-driven, which might be due to the explicit conceptual guidance provided by text to the process of comprehension. Conversely, adaptive mental model specification, as a process of updating the model for specific task requirements, seems to be more (but not exclusively) picture-driven than text-driven, because the nonlinear structure of pictures provides faster and more flexible access to specific information.

Because the text–picture units were randomly selected from authentic secondary school textbooks, the present studies were based on ecologically valid learning material. As compared to the materials used in other studies on multimedia learning (e.g., DeLeeuw & Mayer, 2008), the texts of the present studies were relatively short and the pictures were relatively rich. One could therefore expect that the information-rich pictures supported by frequent text references to the pictures would trigger more picture processing than text processing during initial mental model

Table 4 Study 2: Percentages of accumulated fixation times on text and picture during the no-question phase and the question-answering phase under the single delayed-question condition and during the first and the

second no-question phase and question-answering phase under the twofold delayed-question condition

Variable	(First) No-Question Phase	Second No-Question Phase	Question-Answering Phase
Accumulated Fixation Times: Text/Picture (%)			
Single Delayed Question	76.6%/23.4% (8.6%)	n.a.	47.9%/52.1% (14.49%)
Twofold Delayed Question	72.4%/27.6% (10.2%)	77.6%/22.4% (12.1%)	47.1%/52.9% (17.1%)

construction. In fact, the opposite was found: Participants relied primarily on text for initial mental model construction instead of using pictures, which can be viewed as additional support for the corresponding hypothesis.

This is not to say that pictures were unimportant for initial mental model construction. First, coherence formation with the text–picture units always required both verbal and pictorial information. Second, as Eitel et al. (2013; Eitel & Scheiter, 2015) demonstrated in a series of experiments, pictures allow people to quickly grasp the overall structure of the presented subject matter and serve as a scaffold for initial mental model construction (cf. Lindner et al., 2017). These authors found that even a very short presentation of a picture for a few seconds, without the possibility of an intensive (self-paced) picture analysis, was sufficient to improve comprehension and learning of the subject matter. We also found indications of a scaffolding function of pictures in our time course analyses. However, such a scaffolding function of pictures allows nevertheless for a dominant role of text during initial mental model construction. The findings of Eitel and colleagues therefore do not contradict the results of the present studies.

The goal dependency of cognitively processing written documents, especially text, has been studied by researchers such as Britt et al. (2018), McCrudden and Schraw (2007), Pichert and Anderson (1977), Rickards (1979; Rickards & Denner, 1978), Rouet and Britt (2011), as well as Vidal-Abarca, Mañá, and Gil (2010). Rickards has shown that text processing can be directed toward achieving general goals such as general coherence formation or be directed toward specific goals such as answering certain questions. By presenting different questions before or after the learning material, Rickards demonstrated the possibility of a trade-off between the two kinds of processing—that one goal could be followed at the expense of the other. There is possibly a similar trade-off between general and specific processing, in terms of initial mental model construction and adaptive mental model specification. Initial mental model construction is oriented toward global understanding without a specific task in mind. Adaptive model specification takes place when the learners have a specific task at hand. The latter is highly selective and places special emphasis on task-relevant information. However, there seem to be inherent sequential constraints between initial mental model construction and adaptive model specification, because even highly task-oriented readers seek at least some minimal understanding of what the text and the picture is about. Thus, adaptive mental model specification is assumed to build on preceding initial mental model construction.

The finding that participants who answered many questions correctly spent a larger percentage of fixations on the picture during question answering indicates that

sophisticatedly studying pictures can lead to better performance in answering specific questions. The finding that participants who answered many questions correctly also allocated attention frequently between the text and picture suggests the essential role of text and picture integration during multimedia learning. The effectiveness of teaching and learning in secondary schools might thus be improved by fostering a better understanding of the processes involved in text–picture integration by the teachers, and by providing teachers with guidance in selecting the best remedial strategies to help the students.

The results of these studies indicate that the presentation of instructional material should allow the students to first concentrate on the text in order to receive the required conceptual guidance for the initial mental model construction. This could be achieved by teachers' approach to delivering information as well as by following appropriate guidelines for designing educational materials. As a scaffold for initial mental model construction, an additional pictorial sketch, without superfluous details, might be adequate, as long as no overly detailed pictorial information is presented that could distract the learner's attention from careful reading of the text. After the initial mental model construction, students would be better prepared for an informed and detailed study of pictures for specific purposes.

All in all, our findings support the view that texts and pictures are processed in qualitatively different ways in relationship to the task and have different uses for different purposes, especially for initial mental model construction and adaptive model specification. Texts seem to provide more explicit conceptual guidance in initial mental model construction than do pictures, whereas pictures support mental model adaptation by providing more flexible access to specific information on demand than do texts. Whether the results of the present study can be generalized to other domains or to texts with different contents, different text lengths, or to different kinds of visualizations (Zelazny, 2006) remains to be investigated. It might also be of special interest to study whether and to what extent the kind of text and picture display affects the use of different sources of information in text–picture integration. For example, Sweller and colleagues (Bobis, Sweller, & Cooper, 1994; Chandler & Sweller, 1992) found that cognitive processes can be essentially facilitated by spatially integrating texts and pictures as far as possible. Further research should also use supplementary information sources, such as thinking-aloud data (Hyönä, Radach, & Deubel, 2003; Mason, Pluchino, & Ariasi, 2014), in addition to eye-tracking analysis, to circumvent the ambiguity of eye-tracking indicators. We hope that a deeper understanding of text–picture integration will eventually improve teaching practices to enhance students' competence in studying multimodal documents.

Acknowledgements This study is part of the BITE Project on text–picture integration, funded by the German Research Foundation (Grant Nos. SCHN 665/3-1, SCHN 665/6-1, SCHN 665/6-2) within the Special Research Program “Competence Models for Assessing Individual Learning Results and for Balancing of Educational Processes.” We are grateful to Holger Horz and Mark Ullrich for item development and item analysis during a previous phase of the project.

References

- Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education*, 33, 131–152. doi:[https://doi.org/10.1016/S0360-1315\(99\)00029-9](https://doi.org/10.1016/S0360-1315(99)00029-9)
- Bobis, J., Sweller, J., & Cooper, M. (1994). Demands imposed on primary-school students by geometric models. *Contemporary Educational Psychology*, 19, 108–117. doi:<https://doi.org/10.1006/ceps.1994.1010>
- Britt, M. A., Rouet, J. F., Durik, A. M., Alamargot, D., Chanquoy, L., Albrecht, J. E., . . . Albrecht, J. E. (2018). Situation models in language comprehension and memory. In *Literacy beyond text comprehension: A theory of purposeful reading* (Vol. 21, pp. xiii–xiv). Mahwah, NJ: Erlbaum.
- Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students’ learning from text. *Educational Psychology Review*, 14, 5–26. doi:<https://doi.org/10.1023/A:1013176309260>
- Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, 62, 233–246. doi:<https://doi.org/10.1111/j.2044-8279.1992.tb01017.x>
- Cohn, N. (2016). A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, 146, 304–323. doi:<https://doi.org/10.1016/j.cognition.2015.10.007>
- DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, 100, 223–234. doi:<https://doi.org/10.1037/0022-0663.100.1.223>
- Eitel, A., & Scheiter, K. (2015). Picture or text first? Explaining sequence effects when learning with pictures and text. *Educational Psychology Review*, 27, 153–180. doi:<https://doi.org/10.1007/s10648-014-9264-4>
- Eitel, A., Scheiter, K., Schüler, A., Nyström, M., & Holmqvist, K. (2013). How a picture facilitates the process of learning from text: Evidence for scaffolding. *Learning and Instruction*, 28, 48–63. doi:<https://doi.org/10.1016/j.learninstruc.2013.05.002>
- Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199–241). Cambridge, UK: Cambridge University Press.
- Gyselinck, V., & Tardieu, H. (1999). The role of illustrations in text comprehension: What, when, for whom, and why? In *The construction of mental representations during reading* (pp. 195–218). Mahwah, NJ, US: Erlbaum.
- Hannus, M., & Hyönä, J. (1999). Utilization of illustrations during learning of science textbook passages among low- and high-ability children. *Contemporary Educational Psychology*, 24, 95–123. doi:<https://doi.org/10.1006/ceps.1998.0987>
- Heller, K. A., & Perleth, C. (2000). KFT 4–12+R. Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision [KFT 4–12+R: Cognitive Abilities Test for grades 4 to 12]. Göttingen, Germany: Beltz Test GmbH.
- Hyönä, J., Radach, R., & Deubel, H. (2003). *The mind’s eye: Cognitive and applied aspects of eye movement research*. Amsterdam, The Netherlands: North-Holland.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329–354. doi:<https://doi.org/10.1037/0033-295X.87.4.32>
- Knauff, M., & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & Cognition*, 30, 363–371. doi:<https://doi.org/10.3758/BF03194937>
- Kulhavy, R. W., Lee, J. B., & Caterino, L. C. (1985). Conjoint retention of maps and related discourse. *Contemporary Educational Psychology*, 10, 28–37. doi:[https://doi.org/10.1016/0361-476X\(85\)90003-7](https://doi.org/10.1016/0361-476X(85)90003-7)
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65–100. doi:<https://doi.org/10.1111/j.1551-6708.1987.tb00863.x>
- Lindner, M. A., Eitel, A., Strobel, B., & Köller, O. (2017). Identifying processes underlying the multimedia effect in testing: An eye-movement analysis. *Learning and Instruction*, 47, 91–102. doi:<https://doi.org/10.1016/j.learninstruc.2016.10.007>
- Mason, L., Pluchino, P., & Ariasi, N. (2014). Reading information about a scientific phenomenon on webpages varying for reliability: An eye-movement analysis. *Educational Technology Research and Development*, 62, 663–685. doi:<https://doi.org/10.1007/s11423-014-9356-3>
- Massironi, M. (2002). *The psychology of graphic images: Seeing, drawing, communicating*. Mahwah, NJ: Erlbaum.
- Mayer, R. E. (2009). *Multimedia learning*. Cambridge, UK: Cambridge University Press.
- Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). New York, NY: Cambridge University Press.
- McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review*, 19, 113–139. doi:<https://doi.org/10.1007/s10648-006-9010-7>
- Millis, K. K., & King, A. (2001). Rereading strategically: The influences of comprehension ability and a prior reading on the memory for expository text. *Reading Psychology*, 22, 41–65. doi:<https://doi.org/10.1080/02702710117227>
- Millis, K. K., Simon, S., & tenBroek, N. S. (1998). Resource allocation during the rereading of scientific texts. *Memory & Cognition*, 26, 232–246. doi:<https://doi.org/10.3758/BF03201136>
- Ortega, F. R., Serra, M. J., & England, B. D. (2015). Examining competing hypotheses for the effects of diagrams on recall for text. *Memory & Cognition*, 43, 70–84. doi:<https://doi.org/10.3758/s13421-014-0429-7>
- Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, UK: Oxford University Press, Clarendon Press.
- Pichert, J. W., & Anderson, R. C. (1977). Taking different perspectives on a story. *Journal of Educational Psychology*, 69, 309–315. doi:<https://doi.org/10.1037/0022-0663.69.4.309>
- Rickards, J. P. (1979). Adjunct postquestions in text: A critical review of methods and processes. *Review of Educational Research*, 49, 181–196. doi:<https://doi.org/10.3102/00346543049002181>
- Rickards, J. P., & Denner, P. R. (1978). Inserted questions as aids to reading text. *Instructional Science*, 7, 313–346. doi:<https://doi.org/10.1007/BF00120936>
- Roeschl-Heils, A., Schneider, W., & van Kraayenoord, C. E. (2003). Reading, metacognition and motivation: A follow-up study of German students in grades 7 and 8. *European Journal of Psychology of Education*, 18, 75–86. doi:<https://doi.org/10.1007/BF03173605>
- Rouet, J.-F. (2006). Question answering and document search. In J. F. Rouet (Ed.), *The skills of document use: From text comprehension to web-based learning* (pp. 93–121). Mahwah, NJ: Erlbaum.
- Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). Charlotte, NC: Information Age.
- Sadoski, M., & Paivio, A. (2001). *Imagery and text: A dual coding theory of reading and writing*. Mahwah, NJ: Erlbaum.

- Schnotz, W. (2014). Integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 72–103). Cambridge, UK: Cambridge University Press.
- Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction, 13*, 141–156. doi:[https://doi.org/10.1016/S0959-4752\(02\)00017-8](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*, 850–863. doi:<https://doi.org/10.1037/edu0000246>
- Sims, V. K., & Hegarty, M. (1997). Mental animation in the visuospatial sketchpad: Evidence from dual-task studies. *Memory & Cognition, 25*, 321–332. doi:<https://doi.org/10.3758/BF03211288>
- van der Laan, L. N., Papiés, E. K., Hooge, I. T. C., & Smeets, P. A. M. (2017). Goal-directed visual attention drives health goal priming: An eye-tracking experiment. *Health Psychology, 36*, 82–90. doi:<https://doi.org/10.1037/hea0000410>
- Vidal-Abarca, E., Mañá, A., & Gil, L. (2010). Individual differences for self-regulating task-oriented reading activities. *Journal of Educational Psychology, 102*, 817–826. doi:<https://doi.org/10.1037/a0020062>
- Wainer, H. (1992). *Understanding graphs and tables*. Princeton, NJ: Educational Testing Service.
- Zelazny, G. (2006). *Say it with charts: The executive's guide to visual communication*. Heidelberg, Germany: Redline Wirtschaft.
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.