Applied Cognitive Psychology, *Appl. Cognit. Psychol.* **27**: 451–461 (2013)
Published online 25 March 2013 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.2922

How Inspecting a Picture Affects Processing of Text in Multimedia Learning

ALEXANDER EITEL*, KATHARINA SCHEITER and ANNE SCHÜLER

Knowledge Media Research Center, Tubingen, Germany

Summary: We investigated how a picture fosters learning from text, both with self-paced presentation and with short presentation before text. In an experiment, participants (N = 114) learned about the structure and functioning of a pulley system in one of six conditions: text only, picture presentation for 150 milliseconds, 600 milliseconds, or 2 seconds, or self-paced before text, or self-paced concurrent presentation of text and picture. Presenting the picture for self-paced study time, both before and concurrently with text, fostered recall and comprehension and sped up text processing compared with presenting text only. Moreover, even inspecting the picture for only 600 milliseconds or 2 seconds improved comprehension and yielded faster reading of subsequent text about the spatial structure of the system compared with text only. These findings suggest that pictures, even if attended for a short time only, may yield a spatial mental scaffold that allows for the integration with verbal information, thereby fostering comprehension.

Processing of text and pictures in multimedia learning

In the past three decades, much research has shown that people learn better from text and pictures (i.e., multimedia) than from text alone (see Anglin, Vaez, & Cunningham, 2004; Levie & Lentz, 1982; Mayer, 2009, for reviews). The beneficial effect of learning with a multimedia compared with a mono media message is termed multimedia effect. It constitutes the basis of the cognitive theory of multimedia learning (CTML; Mayer, 2009) as one of the most influential theories in the field of learning with multimedia. According to CTML, pictures foster recall and comprehension of text when they are integrated with text and prior knowledge. Prior to being integrated, relevant words and images are first selected from text and picture, and are then organized into separate mental representations in working memory (i.e., verbal and pictorial mental model). Integration of information from text and pictures is assumed to take place only after separate mental models have been constructed. Thus, CTML does not address whether there is interplay between information extracted from text and pictures at the level of constructing the separate mental representations; that is, it does not comprise any assumptions concerning whether processing of a picture affects processing of text and vice versa. Against the backdrop of research on cognitive psychology, in the present paper, we developed a scaffolding view according to which interplay between text and picture processing is assumed. In particular, according to this view processing of a picture, even for a short time only, is supposed to yield a mental scaffold that facilitates the process of subsequent learning from text.

In research on multimedia learning, students are often required to learn about the structure and functioning of systems comprising cause-and-effect relations in the domain of physics (i.e., causal systems; Mayer & Chandler, 2001). Such causal systems are, for example, bicycle tire pumps (Mayer & Moreno, 2002), hydraulic drum brakes (Mayer, Mathias, & Wetzell, 2002), or pulley systems (Hegarty & Just, 1993). Understanding how these causal systems work (i.e., their functioning) requires knowledge about the system's components and how they are interrelated as well as to be able to mentally animate how the different components of the system move in space and affect each other's movements (Hegarty, Kriz, & Cate, 2003; Narayanan & Hegarty, 2002). This is the case, because understanding how a system works often means that one is able to infer the state of one component of the system given information about the states of the other components and their relations to each other, which is central to being able to design, operate, or troubleshoot the system (Hegarty, 1992).

When learning with text and pictures about how causal systems work, research on multimedia learning has shown that interplay exists in that processing of text affects processing of pictures. This originates from a noteworthy study by Hegarty and Just (1993), in which students learned with the concurrent presentation of text and a picture of a pulley system while their eye movements were recorded. The eye movement data revealed that after reading a sentence or clause, students inspected those parts in the picture about which they had just read in the text. Hegarty and Just concluded that learners first construct an initial understanding of a larger semantic unit in the text (i.e., text base; van Dijk & Kintsch, 1983), before a more elaborate mental model is built that extends this text base with the help of corresponding spatial information extracted from the picture. Learning with text and pictures thus appeared to be largely guided by the text in that the text base directs attention toward those parts of the picture that are addressed in the text. This idea of text-guided processing of pictures is well acknowledged in the literature on learning from text and pictures (e.g., Folker, Ritter, & Sichelschmidt, 2005; Hegarty & Just, 1993; Ozcelik, Arslan-Ari, & Cagiltay, 2010; Rummer, Schweppe, Fürstenberg, Scheiter, & Zindler, 2011; Schmidt-Weigand, Kohnert, & Glowalla, 2010a, 2010b; Schwonke, Berthold, & zRenkl, 2009; Van Gog, Kester, Nievelstein, Giesbers, & Paas, 2009).

On the other hand, to our knowledge less is known on when and how processing of a picture affects processing of text. According to a study by Stone and Glock (1981), there appears to be a picture-initiated processing of text taking place when learning with multimedia. In their study, Stone

*Correspondence to: Alexander Eitel, Knowledge Media Research Center, Schleichstrasse 6, 72076 Tubingen, Germany.
E-mail: a.eitel@iwm-kmrc.de

and Glock asked students to learn about the construction of a loading cart from a text and pictures. Eye movement data revealed that students initially attended to the picture for a short time (i.e., 1 – 2 seconds) prior to processing the text. Stone and Glock concluded that students initially attended to the picture to have a first impression of its overall theme, which they refer to as an attempt to extract the gist from the picture. The question is yet still open whether such an initial glance at the picture already gives enough information about a system's components and their spatial relations to facilitate subsequent learning from text. Alternatively, it could well be that an initial glance at the picture is just an epiphenomenon that occurs arbitrarily when learning from text and pictures without having any functional relevance. The present research seeks to provide answers to these questions. In the following, we argue in favor of the view that an initial glance at a picture does affect subsequent processing of text in a positive manner. In particular, we introduce a scaffolding view according to which spatial information extracted from the picture is assumed to act as mental scaffold, facilitating the process of learning from text—even if the picture has been inspected for a short time only.

Comprehension may benefit from brief initial picture inspection: a scaffolding view

To argue why learning with text may benefit from brief initial picture inspection, in the following we take a closer look at both the nature of information that is extracted from briefly inspecting a picture and the cognitive functions that pictures can play (cf. Ainsworth, 2006; Scaife & Rogers, 1996).

According to theory and research on picture perception, there is reason to assume that a picture's global spatial structure is rapidly extracted, even within a first glance (Navon, 1977; Oliva, 2005). Perception of pictures is thereby assumed to proceed in a global to local manner (cf. Navon, 1977). More specifically, a picture is initially processed as a single entity that is composed of a few global spatial features only (Oliva & Torralba, 2006). To identify a picture's local spatial features, in contrast, a picture needs to be decomposed into its single components, which requires more extensive inspection (Oliva & Torralba, 2006).

Accordingly, studies using pictures of simple geometrical forms have found that the pictures' global spatial structure was initially processed prior to its local spatial structure (Loftus & Harley, 2004; Navon, 1977). Such a global-to-local time course of picture processing has been confirmed in research on the perception of scenes (Castelhano & Henderson, 2007; Henderson & Hollingworth, 1999; Oliva & Schyns, 2000; Rousselet, Joubert, & Fabre-Thorpe, 2005). In research on scene perception, photorealistic pictures of scenes that people experience in their everyday lives are presented to subjects for brief presentation times (e.g., 150 milliseconds; Oliva & Schyns, 2000). Subjects have been shown to be able to categorize scene pictures on a basic semantic level quite accurately from their brief inspection (e.g., Greene & Oliva, 2009). This is known as the ability to quickly identify a scene's gist (e.g., Henderson & Hollingworth, 1999). In contrast, subjects needed substantially more time to accurately identify single objects or details in scene pictures than to identify gist (e.g., Fei-Fei, Iyer, Koch, & Perona, 2007; Liu & Jiang, 2005). Against the backdrop of the spatial envelope model (Oliva & Torralba, 2001, 2006), results for rapid identification of a scene picture's gist followed by slower identification of its details can be interpreted in light of a global-to-local time course of picture processing. Namely initially processing a scene picture as a single entity that is composed of a few global spatial features only (i.e., global spatial structure) was sufficient to infer the scene's gist. In contrast, identification of details in the scene required the scene picture to be decomposed into its local features (i.e., local spatial structure), which required time (Oliva & Torralba, 2006). These assumptions are supported by empirical studies establishing a close relationship between extracting a scene's global spatial structure and extracting its gist (Greene & Oliva, 2009). Hence, findings of rapid extraction of gist followed by slower extraction of details from scene pictures are assumed to reflect a global-to-local time course of picture processing.

Recently, effects of rapid identification of gist followed by slower identification of details have been found with pictures of causal systems that are usually used in studies on learning with multimedia (e.g., a pulley system; Eitel, Scheiter, & Schüler, 2010, 2012). As in scenes, this pattern of results suggests that identification of detail information required the causal system picture to be decomposed into its local spatial structure, which would have required longer inspection times (cf. Oliva & Torralba, 2006). The gist in causal systems, similar to the gist in scenes, thus might have been inferred from the system's global spatial structure that had been extracted from brief inspection of the causal system picture.

The assumption that this information about a system's global spatial structure, in turn, is beneficial for learning from a subsequent text about the causal system is at the heart of the *scaffolding view* on multimedia learning. The scaffolding view is mainly based on two cognitive functions that pictures may play when added to text. First, information about a causal system's spatial structure is much more efficiently extracted from a picture than from a text (Larkin & Simon, 1987). This is the case because in a picture, elements belonging together are grouped together in a meaningful way so that visual search for interrelated elements is minimized (Larkin & Simon, 1987). As a result, learners are able to directly read off a system's spatial structure from the picture, thus reducing the need to process this information in the text. Second, pictures are generally more specific than textual representations in conveying spatial information (Stenning & Oberlander, 1995). Adding a causal system picture to the text may thus resolve ambiguities that are present in text about a causal system's spatial structure, thereby limiting the range of (erroneous) inferences that can be made from the text (graphical constraining; Ainsworth, 2006; Scaife & Rogers, 1996). This may prevent learners from constructing a mental representation that inadequately reflects the spatial structure of the system. As a causal system's functioning is inferred from its spatial structure (e.g., Narayanan & Hegarty, 2002), preventing learners from constructing an inadequate mental representation of the system's spatial structure may result in better comprehension (Schnotz & Bannert, 2003; Schnotz & Kürschner, 2008).

Pictures may exert these facilitative effects not only when presented concurrently with text but also when presented before text as shown by research in the context of the conjoint retention hypothesis (Kulhavy, Lee, & Caterino, 1985; Kulhavy, Stock, & Kealy, 1993). This research has demonstrated that presenting a picture before a corresponding text fosters learning compared with presenting a text before a corresponding picture. This effect is explained by assuming that a spatial representation of a picture can be held as a single unit in working memory while reading a subsequent text without exceeding working memory capacity. The spatial representation of the picture can be assumed to provide a mental scaffold that has beneficial effects on learning from subsequent text (Gyselinck, Jamet, & Dubois, 2008; Kulhavy et al., 1993).

However, in the studies by Kulhavy et al. (1993), pictures were usually presented for as long as learners wanted before the text, thus allowing extraction of the picture's global and local spatial structure. Hence, pictures were inspected for length of times that were much longer than the gist extraction times established in scene perception research or even the times suggested by Stone and Glock (1981). On the basis of the assumption of a global-to-local time course of picture processing (Navon, 1977; Oliva & Torralba, 2006), one may assume that a causal system's global spatial structure only is represented after briefly inspecting the respective picture (i.e., for the time it takes to extract its gist). The causal system's local spatial structure that captures information concerning the system's details, in contrast, is probably not extracted from brief inspection. As spatial information extracted from a picture can act as mental scaffold (Gyselinck et al., 2008), global spatial information extracted from briefly inspecting a picture may act as mental scaffold as well, albeit capturing a system's global spatial structure only. However, a mental scaffold that comprises a causal system's global spatial structure may already constrain interpretation of subsequent text about the system's structure (cf. Scaife & Rogers, 1996), thereby supporting recall and comprehension as well as facilitating processing of text about the system's spatial structure.

Present research and hypotheses

The present research aimed at investigating how processing a picture affects the process of learning from text. Therefore, in an experiment, we investigated how processing a picture both for self-paced study time and for a short time only before text affected learning outcomes and the processing of text about a causal system. According to the scaffolding view developed in the present paper, we assumed interplay between text and picture processing early on in that processing of a picture, even for a short time only, was supposed to facilitate the process of learning from text. In particular, we expected that spatial information that is extracted from both short and self-paced inspection of a causal system picture would be used as a mental scaffold to constrain interpretation of text about the system's spatial structure (cf. Scaife & Rogers, 1996). This should prevent learners from constructing an inadequate mental representation of the system's spatial structure, leading to better results for recall and comprehension (Hypothesis 1). Moreover, spatial information represented within the mental scaffold should facilitate constructing a mental representation of the system's spatial structure from text, reflected in shorter reading times for text about the system's spatial structure (Hypothesis 2).

On the basis of CTML (Mayer, 2009), one may predict better results for recall and comprehension from learning with text and self-paced picture presentation (both before and concurrently with text) than from learning with just text about the causal system. However, on the basis of CTML, one would not predict that processing a picture facilitates processing of text about the causal system. This is the case, because according to CTML, text and pictures are assumed to be first selected and organized into separate mental representations prior to being integrated to have beneficial effects on learning; thus, no interplay is assumed to take place prior to constructing the separate mental representations. As a consequence, briefly inspecting the causal system picture should not be sufficient to already foster learning outcomes, as according to CTML this would require a comprehensive pictorial mental model and not just a partial representation of the system's global spatial features.

EXPERIMENT

The experiment was conducted to assess whether presenting the picture for both self-paced and, more important, for short, system-paced study time in addition to text fosters recall and comprehension, and facilitates the processing of the text about the picture's spatial structure compared with studying the text only. Facilitated processing of text about the picture's spatial structure was operationalized via shorter reading times on the section of the text describing the spatial structure of the pulley system. To test the scaffolding view, we presented the picture of a causal system (i.e., a pulley system; Hegarty & Just, 1993) for brief presentation times prior to the corresponding text. Pictures were presented for 150 , 600 , and 2000 milliseconds, being the presentation times at which gist, but little information about details, can be extracted from causal system pictures (cf. Eitel et al., 2010). As in Kulhavy et al. (1993), we included a condition in which the picture was presented prior to the text, and the picture study time was controlled by the learner (self-paced). Moreover, we included a condition in which text and picture were presented next to each other on one page with self-paced study time, being the usual format of presenting text and pictures in multimedia learning (e.g., Mayer, 2009).

Method

Participants and design

Participants were 114 students (87 female; $M_{age} = 24.25$ years, $SD_{age} = 4.28$; age range: 19 to 54 years) from the University of Tuebingen, Germany, who either were paid 16 euros or received course credit. Students were randomly assigned to one of six experimental conditions (Figure 1). There were 19 participants in each of the six conditions.

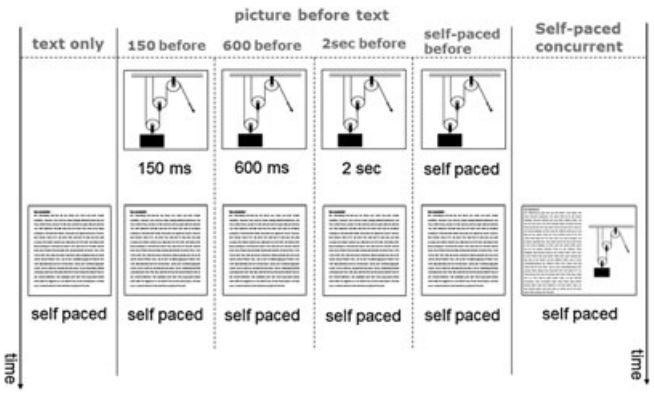

Figure 1. Experimental design. Each column represents an experimental condition

Materials

Participants learned from a text and in five out of six conditions from both a text and a line drawing illustrating the structure and the functioning of pulley systems (Figure 1). Both text and picture were presented on a computer screen. Text and picture fitted on a single page each. In the self-paced concurrent condition, text and picture were presented next to each other on a single page, whereas in the remaining conditions (except for the text-only condition), the picture was presented before the text. The text about pulley systems contained 240 words (Appendix A). In the first section of the text (structure section, 130 words), the spatial structure of the specific pulley system used in the present experiment was described (e.g., 'the upper pulley is attached to the ceiling'). In the second section of the text (functioning section, 110 words), it was described what happens when the rope is pulled (e.g., '. . .middle pulley lifted'; cf. Boucheix & Schneider, 2009; Hegarty & Just, 1993) and how the underlying principles of pulley systems in general work (i.e., each free pulley reduces weight to be lifted by half and doubles the length of rope to be pulled). Note that the text contained information about each component of the pulley system picture so that students were in principle able to recall all single components and perform mental animation appropriately after reading only the text. Information about the principle underlying pulley systems was given only in the text. The picture was a line drawing in black-and-white of the specific pulley system that was described in the text (cf. Hegarty & Just, 1993; Figure 1).

Measures

A short paper-pencil version of the paper folding test (PFT; Ekstrom, French, & Harman, 1976) was administered to test for students' spatial abilities, which have been shown to influence students' ability to mentally animate causal diagrams such as pulley systems (Hegarty, 1992, 2004). To assess prior knowledge, we asked students to draw a picture of a pulley system from memory; that is, they were expected to draw any type of pulley system that they thought of. Therefore, students received a blank sheet of paper with only the instruction 'draw a pulley system' on it. On the basis of the scoring scheme for configuration of pulley systems from Hegarty (1992), one point was assigned for each relation that was drawn in line with the specific pulley system used in the present experiment (e.g., lower pulley attached to weight). This was performed to rule out that participants' mental representations of a pulley system matched the to-be-learned pulley system in the experiment, indicating that they had possibly previously seen this particular pulley system. There were a total of 11 relations that could have been drawn correctly, and thus the maximum score for prior knowledge was 11 points. Two independent raters scored the drawings from 30 of the 114 participants (26%). Inter-rater agreement was sufficiently high ($r = .91$, $p < .001$) so that only one rater continued scoring the drawings of the remaining participants.

Students' learning outcomes were measured by means of recall (pictorial and verbal) and comprehension items. Pictorial recall was assessed by student's drawings of pulley systems after learning from text only or text and picture, respectively. These drawings were scored in the same way as the drawings that students had generated prior to learning in that one point was assigned to each relation that was drawn in line with the to-be-learned pulley system. Inter-rater agreement on 26% of the data was sufficiently high ($r = .97$, $p < .001$) so that only one rater continued scoring the drawings of the remaining participants.

Verbal recall was tested with eight items in a verbal multiple-choice format, where the students had to judge the correctness of statements such as 'both ropes are attached to the ceiling with one end' by checking 'yes' or 'no'. Cronbach's alpha for this test was .69. Comprehension of pulley systems was assessed in terms of students' ability to mentally animate the system (Hegarty & Just, 1993) and to understand the (abstract) principle underlying pulley systems (i.e., each free pulley splits weight to be lifted in half). Being able to mentally animate the system was assessed via nine items in a verbal multiple-choice format, where the students had to verify whether statements such as 'if the free end of the upper rope is let go, then the middle pulley turns clockwise' was correct. Being able to understand the principle underlying pulley systems was assessed via three items in a verbal multiple-choice format, where the students had to judge the correctness of statements such as 'if the weight was attached to the middle pulley, then the rope would have to be pulled with the same force as when the weight is attached to the lower pulley' and via four items in a labeling test. In this labeling test, students were asked to indicate how much the weight to-be-lifted is reduced in a depiction of one specific pulley system compared with another one. Depictions of pulley systems differed in the number of free pulleys so that correctly solving the task required comprehension of the principle of free pulleys. Results from the verbal multiple-choice items and the labeling items were merged in the analysis so that the comprehension test consisted of 16 items in total. As each correct response was credited one point and each incorrect response was given zero point, students could score a minimum score of 0 and a maximum score of 16 points in the comprehension test about pulley systems. Cronbach's alpha for this test was .58. For each item in the multiple-choice test, students had to rate how confident they felt with regard to their response on a Likert-type scale ranging from 'guessed' to 'surely known' (cf. Cierniak, Scheiter, & Gerjets, 2009). If students marked the lowest score in their

confidence rating ('guessed'), the respective response was multiplied by 0 so that guessed responses were not counted in the analysis. All responses with a confidence rating that was higher than 'guessed' were counted in the analysis (multiplied by 1). This was performed to bypass the problem of guessing probability leading to an increased reliability of the multiple-choice tests used in the present experiments (cf. Cierniak et al., 2009; Conway, Gardiner, Perfect, & Cohen, 1997).

We assessed reading times for the section of the text describing the spatial structure of the pulley system (structure section) and for the section of the text explaining how pulley systems work (functioning section) by means of eye tracking. To do so, areas of interest (AoIs) were drawn around both the structure and the functioning section in the text, and the dwell times that learners spent on the two text sections (i.e., reading times) were computed.

Procedure

Students were tested in single sessions of approximately 50 minutes. Students were first given a demographic questionnaire followed by the PFT and the prior knowledge test. Students were then seated in front of a computer screen. Prior to presenting the text and picture about pulley systems, a text and a picture about a toilet flush were presented (without the instruction to learn) so that students could familiarize themselves with the experimental procedure. After this training trial, the text and picture about pulley systems were presented, and students were instructed to acquire as much information as possible from the multimedia instruction. Its presentation was preceded by a fixation cross that was displayed for 800 milliseconds so that students could prepare for the upcoming presentation of text and picture. In conditions with picture-before text, the picture appeared on the screen for 150 milliseconds, 600 milliseconds or 2 seconds, or the picture stayed on the screen until students signaled that they had sufficiently inspected the picture (self-paced). The experimenter responded to the signal by pressing a key so that the presentation of the picture was replaced by a mask that was displayed for 500 milliseconds. In the text-only condition, the text appeared right after the fixation cross. Reading the text was self-paced in all experimental conditions. In the self-paced concurrent condition, text and picture appeared right after the fixation cross, and learning was self-paced as well.

After learning about pulley systems, students were again instructed to draw a picture of the pulley system about which they had just learned. Subsequently, students were given the verbal multiple-choice items and the labeling test.

Apparatus

During learning, eye movements were recorded with a video-based eye tracking system (iView X™ Hi-Speed 1250) from SensoMotoric Instruments (SMI, Teltow, Germany) with a 500-Hz sampling rate. The system was calibrated using a 13-point calibration image. Stimuli were presented using E-prime 2.0 Professional from Psychology Software Tools® (Sharpsburg, Pennsylvania). Eye tracking data were recorded using iView X™ from SMI.

Results

First, we analyzed students' spatial abilities and prior knowledge to determine whether the six instructional conditions were comparable with regard to the students' prior abilities. Then learning outcome measures (verbal and pictorial recall, and comprehension) and text-processing measures (dwell times on the two sections) were analyzed as a function of experimental condition by means of analysis of covariance (ANCOVA). Subsequent planned comparisons were conducted to test Hypotheses 1 and 2. Accordingly, planned comparisons were expected to yield significant differences between the text-only condition and the conditions with both brief initial (i.e., 150 milliseconds, 600 milliseconds and 2 seconds) and self-paced picture inspection (i.e., self-paced before and self-paced concurrent) regarding verbal and pictorial recall, comprehension, and reading time on the section of the text describing the system's spatial structure. Last, correlations between picture-inspection times, text-processing times, and learning outcomes were analyzed in conditions with self-paced picture inspection.

Prior abilities

Descriptive values are shown in Table 1. A one-factorial analysis of variance revealed that students' spatial abilities differed between experimental conditions, $F(5, 108) = 3.10$, $MSE = 3.45$, $p = .01$, $\eta_p^2 = 0.13$. In addition, spatial abilities

Table 1. Descriptive data for prior abilities and learning outcomes as a function of experimental condition

	Text only	150 milliseconds before	600 milliseconds before	2 seconds before	Self-paced before	Self-paced concurrent	*M*
Spatial abilities (min. = 0, max. = 10)	6.16 (1.68)	7.00 (1.83)	5.56 (1.73)	6.90 (2.16)	6.97 (2.17)	5.32 (1.49)	6.32 (1.94)
Prior knowledge (min. = 0, max. = 11)	0.11 (0.32)	0.16 (0.38)	0.37 (0.83)	0.26 (0.65)	0.21 (0.42)	0.05 (0.23)	0.19 (0.51)
Pictorial recall (min. = 0, max. = 11)	6.11 (0.86)	5.21 (0.87)	6.18 (0.87)	5.99 (0.86)	8.75 (0.87)	9.86 (0.88)	7.02 (0.35)
Verbal recall (min. = 0, max. = 8)	6.09 (0.41)	5.29 (0.41)	6.20 (0.42)	5.53 (0.41)	6.61 (0.41)	6.84 (0.42)	6.10 (0.17)
Comprehension (min. = 0, max. = 16)	8.33 (0.55)	8.16 (0.56)	9.97 (0.56)	9.91 (0.56)	11.55 (0.56)	11.67 (0.56)	9.93 (0.23)

Note: Means and standard deviations for spatial abilities and prior knowledge as well as the adjusted marginal means (and standard errors) corrected for the influence of spatial abilities are reported.

were correlated to some of the dependent variables in the current experiment (Table 2), which is why they were included as a covariate in the statistical analyses. Prior knowledge levels of students as indicated by their drawings of a pulley system from memory prior to learning were low, suggesting that the students were not familiar with the particular system, and did not differ between conditions, $F < 1$.

Recall

Descriptive values are shown in Table 1. Two ANCOVAs with condition as factor (text only vs. 150 milliseconds before vs. 600 milliseconds before vs. 2 seconds before vs. self-paced before vs. self-paced concurrent), spatial abilities as covariate, and verbal and pictorial recall for pulley systems as dependent variables were conducted.

The ANCOVA for verbal recall failed to reveal a significant main effect of condition, $F(5, 107) = 2.05$, $MSE = 3.18$, $p = .08$, $\eta_p^2 = 0.09$. Also the planned comparisons revealed that verbal recall did not significantly differ between any of the conditions with picture presentation compared with the text-only condition (all ps $> .15$). Regarding pictorial recall, there was a significant main effect of condition, $F(5, 107) = 4.52$, $MSE = 13.91$, $p = .001$, $\eta_p^2 = 0.17$. Planned comparisons revealed that presenting the picture for a short time before the text (150 milliseconds, 600 milliseconds and 2 seconds) did not foster pictorial recall compared with presenting the text only (all ps $> .40$). On the other hand, presenting the picture for self-paced, both before ($p = .03$) and concurrently with the text ($p = .003$), fostered pictorial recall compared with the text-only condition. Furthermore, polynomial contrasts were conducted to assess how scores for pictorial recall developed as a function of experimental condition. Polynomial contrasts showed a significant linear trend ($p < .001$) as well as a significant quadratic trend ($p = .03$). As no further trend was significant (all ps $> .25$), this pattern of results suggests that scores for pictorial recall did not differ between the first conditions, but increased in the later conditions.

Comprehension

Descriptive values are shown in Table 1. An ANCOVA with condition as independent variable, spatial abilities as covariate, and comprehension for pulley systems as dependent variable revealed a significant main effect of condition, $F(5, 107) = 7.40$, $MSE = 5.75$, $p < .001$, $\eta_p^2 = 0.26$ (Figure 2). Planned comparisons showed that comprehension scores did not differ between the text-only condition and the 150-milliseconds-before condition ($p = .91$). By contrast, comprehension was better in the 600-milliseconds-before ($p = .04$), in the 2-seconds-before ($p = .046$), in the self-paced before ($p < .001$), and in the self-paced concurrent condition ($p < .001$) compared with the text-only condition. In addition, polynomial contrasts revealed a significant linear trend ($p < .001$). As no further trend was significant (all ps .10), this result suggests that comprehension scores increased linearly as a function of experimental condition.

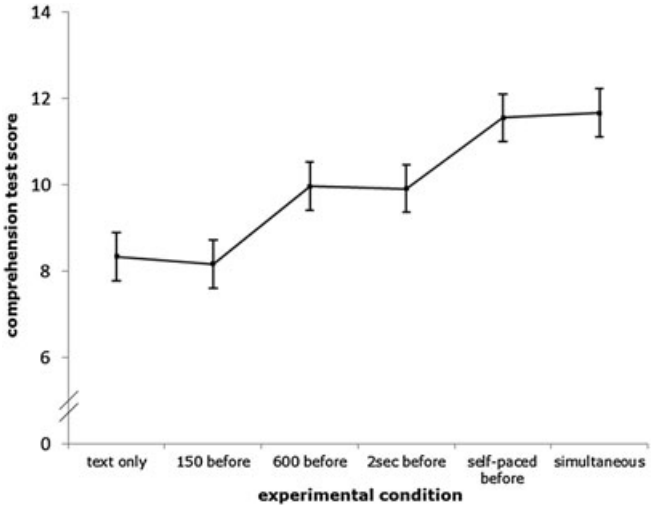

Figure 2. Adjusted means and standard errors for comprehension as a function of experimental condition

Text processing

Descriptive values for reading times (dwell times on AoIs) are shown in Table 3. To investigate effects of the experimental manipulation on the processing of the texts about the structure and the functioning of the pulley system separately, we introduced text type (structure section vs. functioning section) as a repeated-measures factor in the ANCOVA. As a consequence, we conducted a 6×2 ANCOVA with condition (text only vs. 150 milliseconds before vs. 600 milliseconds before vs. 2 seconds before vs. self-paced before vs. self-paced concurrent) as between-subjects factor, text type (structure section vs. functioning section) as within-subjects factor, spatial abilities as covariate, and reading time as dependent variable. The ANCOVA revealed significant main effects of condition, $F(5, 107) = 6.29$, $MSE = 1491.76$, $p < .001$, $\eta_p^2 = 0.23$, and text type, $F(1, 107) = 25.91$, $MSE = 1305.02$, $p < .001$, $\eta_p^2 = 0.20$. Most important, the ANCOVA revealed a significant interaction between both factors, $F(5, 107) = 6.22$, $MSE = 1305.02$, $p < .001$, $\eta_p^2 = 0.23$ (Figure 3).

To break down the interaction, separate ANCOVAs were conducted for the structure and the functioning section. The ANCOVA for the structure section revealed a significant main effect of condition, $F(5, 107) = 7.24$, $MSE = 3550.34$, $p < .001$, $\eta_p^2 = 0.25$, whereas the ANCOVA for the functioning

Table 2. Correlations between the unequally distributed spatial abilities of participants and the dependent variables in the experiment ($N = 114$)

	Verbal recall	Pictorial recall	Comprehension	Learning time	Reading time	Reading time on structure section	Reading time on functioning section
Spatial abilities	$r = .21$*	$r = .11$	$r = .36$**	$r = -.19$*	$r = -.16$	$r = -.14$	$r = -.16$
	$p = .02$	$p = .27$	$p < .001$	$p = .05$	$p = .09$	$p = .13$	$p = .09$

Note: *$p < .05$; **$p < .001$.

Table 3. Descriptive data for picture inspection and reading times (dwell times in seconds) as a function of experimental condition

	Text only	150 milliseconds before	600 milliseconds before	2 seconds before	Self-paced before	Self-paced concurrent	*M*
Picture inspection time	0 (0)	0.15 (0)	0.60 (0)	2.00 (0)	30.54 (3.96)	30.34 (3.96)	—
Reading time on structure section	155.02 (13.68)	136.75 (13.83)	114.43 (13.87)	108.05 (13.79)	84.89 (13.82)	50.90 (14.01)	108.34 (5.58)
Reading time on functioning section	56.69 (6.24)	55.36 (6.31)	49.80 (6.32)	47.98 (6.29)	41.24 (6.30)	37.08 (6.39)	48.03 (2.55)
Overall reading time	211.71 (17.73)	192.11 (17.93)	164.23 (17.98)	156.03 (17.87)	126.14 (17.91)	87.97 (18.17)	156.36 (7.24)

Note: The adjusted marginal means (and standard errors) corrected for the influence of spatial abilities are reported.

section did not, $F(5, 107) = 1.52$, $MSE = 738.20$, $p = .19$, $\eta_p^2 = 0.07$. In consequence, planned comparisons were conducted only for the structure section. They revealed that reading times of students who saw the picture for 150 milliseconds before the text did not differ from students in the text-only condition ($p = .35$). By contrast, they revealed that students in the 600-milliseconds-before ($p = .04$), in the 2-seconds-before ($p = .02$), in the self-paced before ($p < .001$), and in the self-paced concurrent condition ($p < .001$) had shorter reading times than students in the text-only condition. In addition, polynomial contrasts revealed a significant linear trend ($p < .001$). As no further trend was significant (all ps $> .40$), this result suggests that reading times on the structure section of the text decreased linearly as a function of experimental condition.

Links between picture-inspection times, text-processing times, and performance

Descriptive values for picture-inspection times are shown in Table 3. How picture-inspection times, text-processing times, and learning outcomes were related to each other in conditions with self-paced picture inspection (i.e., self-paced before and self-paced concurrent) is analyzed in the following by means of correlations ($N = 38$). Picture-inspection times were positively correlated to the time spent reading the text ($r = .43$, $p = .008$). Furthermore, picture-inspection times were positively correlated pictorial recall ($r = .37$, $p = .02$); however, there were no significant correlations between picture-inspection times and verbal recall ($r = .28$, $p = .09$) or comprehension ($r = .09$, $p = .60$). Text-processing times were not correlated to verbal recall ($r = .25$, $p = .13$) nor to pictorial recall ($r = .14$, $p = .41$) nor to comprehension ($r = .09$, $p = .58$).

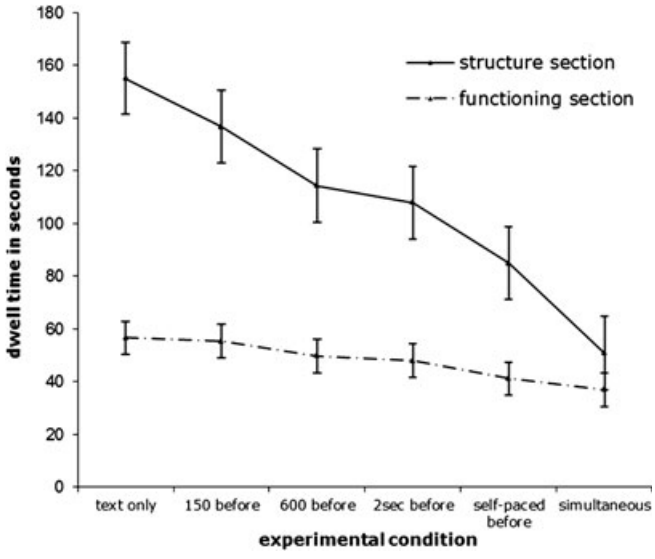

Figure 3. Adjusted means and standard errors for reading times about the spatial structure and about the functioning of pulley systems (dwell times on areas of interest) as a function of experimental condition

DISCUSSION

The present research aimed at investigating how processing a picture affects the process of learning from text. Hypotheses were derived from a scaffolding view on multimedia learning, according to which early interplay between processing of pictures and processing of text was assumed; that is, presenting a picture was assumed to have beneficial effects on the process of learning from text—even if the picture was inspected for a short time only.

According to Hypothesis 1, inspecting the causal system picture both for self-paced study time (before and concurrently with text) and for a short time before text was supposed to foster verbal and pictorial recall and comprehension compared with learning with text only. In support of Hypothesis 1, presenting the picture for self-paced study time both before and concurrently with the text fostered pictorial recall and comprehension; moreover, even presenting the picture for a short time before text (i.e., for 600 milliseconds and 2 seconds) fostered comprehension compared with learning with text only. On the basis of the scaffolding view, we explain these results by assuming that spatial information extracted from both self-paced and brief initial picture inspection was used as mental scaffold to constrain interpretation of text about the system's spatial structure (cf. Scaife & Rogers, 1996). This prevented learners from constructing an inadequate mental representation of the system's spatial structure, resulting in better comprehension than when learning with just text.

According to Hypothesis 2, inspecting the causal system picture both for self-paced study time (before and concurrently with text) and for a short time before text was supposed to facilitate processing of text as reflected in shorter reading times on text about the system's spatial structure compared with learning with text only. In support of Hypothesis 2, results revealed that students who saw a picture both for self-paced study time (i.e., before and concurrently with text) and for a short time before text (i.e., for 600 milliseconds and 2 seconds) processed text about the causal system's spatial structure faster than students who learned with text only. According to the scaffolding view,

these findings are explained by assuming that the spatial information represented within the mental scaffold facilitated constructing a mental representation of the system's spatial structure from text. This was reflected in shorter reading times on text about the system's spatial structure.

As processing of text about the system's spatial structure but not about its functioning was sped up by the picture, one may conclude that the mental scaffold extracted from the picture comprised the system's (global) spatial structure without much information concerning its functioning. This notion is further supported by the result that comprehension levels increased linearly with increasing picture-inspection times along the six experimental conditions. Namely this shows that the longer the picture was inspected the better the comprehension level, suggesting that the more spatial information from the picture was represented within the mental scaffold, the better it constrained interpretation of text. The better the mental scaffold constrained interpretation of text, the more it prevented learners from constructing a mental representation from text that inadequately reflects the spatial structure of the system in turn, thereby fostering comprehension (Schnotz & Bannert, 2003).

Furthermore, results revealed that the time students took to process text about the system's spatial structure decreased linearly with increasing picture-inspection times along the six experimental conditions. These results indicate that there was interplay between processing of the picture and processing of text in the present experiment. In particular, they show that the more information concerning the system's spatial structure was processed in the picture, the less spatial information was needed to be processed in text. However, as pictures provide direct and rapid access to specific spatial information (cf. Larkin & Simon, 1987; Oliva & Torralba, 2006; Stenning & Oberlander, 1995), the spatial information extracted from a first glance at the picture was already sufficient to facilitate the process of learning from text. Taken together, these results suggest that spatial information processed in the picture might have facilitated the process of constructing a mental representation of the system's spatial structure from text, even though only global spatial information was extracted from brief initial picture inspection. This indicates interplay between processing of the picture and processing of text already at an early stage when learning from both media.

In contrast, CTML (Mayer, 2009) does not address such interplay between processing of text and processing of pictures when learning with multimedia. As a result, on the basis of CTML, one would not assume that presenting a picture in addition to text would influence the processing of text in multimedia learning, as revealed by the present results. Rather, according to CTML, text and pictures are assumed to be processed in separate channels in working memory; integration of information from text and pictures is assumed to take place only after separate mental models have been constructed. Thus, results of better comprehension from presenting a picture for a short time only before text might not have been explained by CTML either, that is because brief initial inspection of the causal system picture presumably was not sufficient to construct a comprehensive pictorial mental model, but rather to just construct a partial representation of the system's global spatial features (cf. Eitel et al., 2010; Greene & Oliva, 2009). Only a self-paced picture study time might have been long enough to allow constructing a comprehensive pictorial mental model that could then be integrated with the verbal mental model constructed from text to foster recall and comprehension. Importantly, our findings do not stand in conflict with CTML; rather, they refer to aspects that have not yet been focused on within the theory. Accordingly, once our findings have been replicated potentially with a wider range of materials, they would allow expanding CTML's assumptions concerning the processing of text and pictures at a more fine-grained level by addressing the interplay that may occur between these two processes.

Before such an extension is warranted, however, some limitations have to be ruled out that occurred for the results obtained within the present experiment. For instance, as the internal consistency of the comprehension test was rather low, one has to treat results for better comprehension with caution. This is, however, explainable in the case of knowledge tests, because multiple items are used to assess different and partly independent aspects of the learning domain. For example, students may be able to mentally animate their mental representations of a pulley system, whereas they may fail to understand the principle of free pulleys that underlies the functioning of pulley systems. This pattern of answering would yield low internal consistency values regardless of the quality of the measurement.

Moreover, only six items in a multiple-choice format were used to measure verbal recall. Thus, the fact that presenting the picture neither for self-paced study time nor for a short time before text fostered verbal recall compared with learning from text only may go back to a low power of the verbal recall test used in the experiment. Besides, the verbal recall test might not have been very sensitive regarding the experimental manipulations.

In addition, different than expected, presenting the picture for only 150 milliseconds before the text did not have any beneficial effects on comprehension and text processing compared with presenting the text alone. It may be that the spatial representation extracted from a 150-millisecond inspection of the causal system picture was not sufficiently stable or reliable to affect subsequent reading and comprehension processes, unlike the spatial representation extracted from 600-millisecond and 2-second inspection of the picture. Moreover, presenting the picture for a short time (i.e., for 150 milliseconds, 600 milliseconds, and 2 seconds) did not foster pictorial recall. One might explain this missing effect by taking the nature of the hypothesized mental scaffold into account; namely representing the pulley system's global spatial structure within the mental scaffold was sufficient to foster comprehension but not sufficient to foster recall as measured in the present experiment. Here pictorial recall was measured predominantly in terms of how the single objects in the pulley system (e.g., pulleys) are connected to each other via the ropes. Such relations can be considered as being part of the system's *local* spatial structure. Presumably, the system's local spatial structure was not extracted from brief inspection (cf. Eitel et al., 2010). Instead, extracting the system's local spatial structure to foster pictorial recall required decomposition of the picture into its local

spatial features, in turn requiring a more extensive picture inspection (cf. Oliva & Torralba, 2006). This may elucidate why we did not find beneficial effects of brief initial picture presentation on performance in the pictorial recall test in the present experiment. However, one may assume that brief initial picture presentation might yield beneficial effects if recall of the system's global spatial structure is measured. This should be subject to further studies.

Similarly, results from the correlational analysis in the conditions with self-paced picture inspection may be explained in light of the distinction between extraction of the system's global and local spatial structure. Namely pictorial recall improved with more intensive studying of the picture, suggesting that the process of decomposing the picture into its local spatial features had been performed more thoroughly at prolonged picture-inspection times. As the pictorial recall test asked for information concerning the system's local spatial structure, prolonged picture inspection that presumably reflects more thorough picture decomposition was related to better performance in this test.

Moreover, there was a positive correlation between picture study time and the time for processing the text in the present experiment. Hence, longer picture study times were related to more thorough reading. One may explain this finding by taking into account general differences in students' processing styles (cf. Gernsbacher, Varner, & Faust, 1990). Students who take longer to process a picture may also take longer to process respective text and vice versa. Picture study times are thus a priori positively correlated with the time to process text. This is reflected in the results of the present experiment.

It is interesting that learning outcomes were on a similar level in the condition with self-paced picture inspection before text compared with the self-paced concurrent condition. At first glance, this may contradict the temporal contiguity principle on multimedia learning according to which text and picture should be presented concurrently rather than sequentially to facilitate integration (cf. for reviews Ginns, 2006; Mayer, 2009). However, this principle is limited to the use of spoken texts, whereas no negative effects of sequential presentation are typically found when using written texts (Michas & Berry, 2000). On the contrary, when using written texts, the temporal contiguity effect may even be reversed, thereby speaking in favor of a sequential representation (Rummer et al., 2011; Schüler, Scheiter, Rummer, & Gerjets, 2012).

In the present experiment, beneficial effects from adding a picture to text were obtained in a sequential *picture-before-text* presentation format. Beneficial effects were hypothesized to occur, because pictures provide a computational advantage when processed before text (Kulhavy et al., 1993; Schnotz, 2005); that is, pictures provide more direct and effortless access to accurate information about a causal system's spatial structure than text (Larkin & Simon, 1987; Stenning & Oberlander, 1995). Moreover, the spatial representation of a picture can be held active in working memory while processing text, thus allowing for its integration with text (Kulhavy et al., 1993). As a result, the processing of a text about a causal system is supported by spatial information extracted and integrated from previous picture inspection that, in turn, may foster comprehension (Mayer et al., 2002; Narayanan & Hegarty, 2002).

The computational advantage of pictures may be lost when they are presented after text (Kulhavy et al., 1993; Schnotz, 2005), in which case a mental representation of the system's spatial structure has to be first constructed from text. As a text is usually more ambiguous than a picture, it has to be interpreted to construct a mental representation of the system's specific spatial structure (cf. Ainsworth, 2006; Stenning & Oberlander, 1995). This may lead to the construction of a mental representation about the system's spatial structure that deviates from how the spatial structure is illustrated in a subsequently presented picture. Thus, the two representations will likely interfere, possibly hampering comprehension (Schnotz, 2005). Accordingly, empirical studies revealed more successful learning when a picture had been presented before rather than after a corresponding text (Baggett, 1984; Kulhavy et al., 1993; Ullrich & Schnotz, 2008).

In the present study, in four out of six conditions, pictures were presented prior to text about the structure and functioning of pulley systems. Presenting the picture before the text may share similarities with a pre-training phase, during which students are familiarized with the names and characteristics of the key parts of the system to-be-learned. Pre-trainings have been shown to have beneficial effects on learning (Clarke, Ayres, & Sweller, 2005; Mayer et al., 2002). However, different from pre-trainings, scaffolding fosters learning by (rapidly) providing spatial information that helps to construct a mental representation from text rather than introducing single components in isolation. Thus, the function of presenting a picture prior to text is more specific than that of pre-trainings.

Finally, it remains an open question whether students take a glance at the picture spontaneously prior to processing text in multimedia learning. To our knowledge to date, there exists only one study in the context of multimedia learning, providing tentative evidence in favor of the view that students do take a brief initial glance at the picture by themselves (Stone & Glock, 1981). However, several studies in other research contexts have shown that humans do take a brief initial glance at the picture prior to reading a text, for instance, when processing advertisements (Rayner, Miller, & Rotello, 2008; Rayner, Rotello, Stewart, Keir, & Duffy, 2001), comics (Carroll, Young, & Guertin, 1991), or real-world scenes (Underwood, Jebbett, & Roberts, 2004). Further research in the context of multimedia learning is needed to determine which entry point to processing text–picture combinations students typically take (i.e., text or picture). A study by Scheiter and Eitel (2010) showed that learners more frequently used the picture as an entry point to processing a text–picture combination when visual signals were added to text and picture. However, the question of what medium is processed first will most likely also depend on several other features of text and pictures that can be quickly accessed by learners such as text length, salience of the picture, or of single components.

Regardless of whether picture or text is habitually processed first, results from the present studies show that an initial glance at the picture can be helpful to learning from a corresponding text. As a consequence, results from the present studies provide first tentative evidence in favor of a scaffolding view on multimedia learning, namely that a

mental scaffold is extracted from the initial glance at the picture, which, in turn, facilitates construction of an adequate mental representation from subsequent text.

ACKNOWLEDGEMENTS

This research is funded by the Pact for Research and Innovation of the Competition Fund of the Leibniz Gemeinschaft. We would like to thank Paul Ayres, Krista DeLeeuw, Erlijn van Genuchten, Kim Stalbovs, and two anonymous reviewers for valuable comments on earlier versions of this article. We also thank Elisabeth Arzberger for the help with data collection.

REFERENCES

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*, 183–198. doi:10.1016/j.learninstruc.2006.03.001

Anglin, G. J., Vaez, H., & Cunningham, K. L. (2004). Visual representations and learning: The role of static and animated graphics. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 865–916). Mahwah, NJ: Lawrence Erlbaum.

Baggett, P. (1984). Role of temporal overlap of visual and auditory material in forming dual media associations. *Journal of Educational Psychology*, *76*, 408–417. doi:10.1037//0022-0663.76.3.408

Boucheix, J.-M., & Schneider, E. (2009). Static and animated presentations in learning dynamic mechanical systems. *Learning and Instruction*, *19*, 112–127. doi:10.1016/j.learninstruc.2008.03.004

Carroll, P. J., Young, J. R., & Guertin, M. S. (1991). Visual analysis of cartoons: A view from the far side. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 444–461). New York: Springer.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology. Human Perception and Performance*, *33*, 753–763. doi:10.1037/0096-1523.33.4.753

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior*, *25*(2), 315–324. doi:10.1016/j.chb.2008.12.020

Clarke, T., Ayres, P., & Sweller, J. (2005). The impact of sequencing and prior knowledge on learning mathematics through spreadsheet applications. *Educational Technology Research and Development*, *53*, 15–24. doi:10.1007/BF02504794

Conway, M. A., Gardiner, J. M., Perfect, T. J., & Cohen, G. M. (1997). Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology. General*, *126*, 393–413. doi:10.1037/0096-3445.126.4.393

Eitel, A., Scheiter, K., & Schüler, A. (2010). What can information extraction from scenes and causal systems tell us about learning from text and pictures? In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2822–2827). Austin, TX: Cognitive Science Society.

Eitel, A., Scheiter, K., & Schüler, A. (2012). The time course of information extraction from instructional diagrams. *Perceptual and Motor Skills*, *155*, 677–701. doi:10.2466/22.23.PMS.115.6.677-701

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton: Educational Testing Service.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P.(2007). What do we perceive in a glance at a real-world scene? *Journal of Vision*, *7*, 1–29. doi:10.1167/7.1.10

Folker, S., Ritter, H., & Sichelschmidt, L. (2005). Processing and integrating multimodal material: The influence of color-coding. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the cognitive science society* (pp. 690–695). Mahwah, NJ: Erlbaum.

Gernsbacher, M. A., Varner, K. R., & Faust, M. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 430–445.

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

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*, 137–176. doi:10.1016/j.cogpsych.2008.06.001

Gyselinck, V., Jamet, E., & Dubois, V. (2008). The role of working memory components in multimedia comprehension. *Applied Cognitive Psychology*, *22*, 353–374. doi:10.1002/acp.1411

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1084–1102. doi:10.1037/0278-7393.18.5.1084

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*, 280–285. doi:10.1016/S1364-6613(04)00100-7

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

Hegarty, M., Kriz, S., & Cate, C. (2003). The roles of mental animations and external animations in understanding mechanical systems. *Cognition and Instruction*, *21*, 325–360. doi:10.1207/s1532690xci2104_1

Henderson, J. M., & Hollingworth, A. (1999). High level scene perception. *Annual Review of Psychology*, *50*, 243–271. doi:10.1146/annurev.psych.50.1.243

Kulhavy, R. W., Lee, J. B., & Caterino, L. C. (1985). Conjoint retention of maps and related discourse. *Contemporary Educational Psychology*, *10*, 28–37. doi:10.1016/0361-476X(85)90003-7

Kulhavy, R. W., Stock, W. A., & Kealy, W. A. (1993). How geographic maps increase recall of instructional text. *Educational Technology Research and Development*, *41*, 47–62. doi:10.1007/BF02297511

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65–99. doi:10.1016/S0364-0213(87)80026-5

Levie, W. H., & Lentz, R. (1982). Effects of text illustrations: A review of research. *Educational Communication and Technology*, *30*, 195–232.

Liu, K., & Jiang, Y. (2005). Visual working memory for briefly presented scenes. *Journal of Vision*, *5*, 650–658. doi:10.1167/5.7.5

Loftus, G. R., & Harley, E. M. (2004). How different spatial-frequency components contribute to visual information acquisition. *Journal of Experimental Psychology. Human Perception and Performance*, *30*, 104–118. doi:10.1037/0096-1523.30.1.104

Mayer, R. E. (2009). *Multimedia learning*, 2nd edition. Cambridge: Cambridge University Press.

Mayer, R. E., & Chandler, P. (2001). When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *Journal of Educational Psychology*, *93*, 390–397. doi:10.1037//0022-0663.93.2.390

Mayer, R. E., & Moreno, R. (2002). Aids to computer-based multimedia learning. *Learning and Instruction*, *12*, 107–119. doi:10.1016/S0959-4752(01)00018-4

Mayer, R. E., Mathias, A., & Wetzell, K. (2002). Fostering understanding of multimedia messages through pre-training: Evidence for a two-stage theory of mental model construction. *Journal of Experimental Psychology. Applied*, *8*, 147–154. doi:10.1037//1076-898X.8.3.147

Michas, I. C., & Berry, D. C. (2000). Learning a procedural task: Effectiveness of multimedia presentations. *Applied Cognitive Psychology*, *14*, 555–575. doi:10.1002/1099-0720(200011/12)14:6<555::AID-ACP677>3.3.CO;2-W

Narayanan, N. H., & Hegarty, M. (2002). Multimedia design for communication of dynamic information. *International Journal of Human Computer Studies*, *57*, 279–315. doi:10.1006/ijhc.1019

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353–383. doi:10.1016/0010-0285(77)90012-3

Oliva, A. (2005). Gist of the Scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.). *Neurobiology of attention* (pp. 251–256). San Diego, CA: Elsevier.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, *41*, 176–210. doi:10.1006/cogp.1999.0728

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36. doi:10.1016/S0079-6123(06)55002-2

Ozcelik, E., Arslan-Ari, I., & Cagiltay, K. (2010). Why does signaling enhance multimedia learning? Evidence from eye movements. *Computers in Human Behavior*, *26*, 110–117. doi:10.1016/j.chb.2009.09.001

Rayner, K., Miller, B., & Rotello, C. M. (2008). Eye movements when looking at print advertisements: The goal of the viewer matters. *Applied Cognitive Psychology*, *22*, 697–707. doi:10.1002/acp.1389

Rayner, K., Rotello, C. M., Stewart, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied*, *7*, 219–226. doi:10.1037//1076-898X.7.3.219

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*, 852–877. doi:10.1080/13506280444000553

Rummer, R., Schweppe, J., Fürstenberg, A., Scheiter, K., & Zindler, A. (2011). The perceptual basis of the modality effect in multimedia learning. *Journal of Experimental Psychology: Applied*, *17*, 159–173. doi:10.1037/a0023588

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human Computer Studies*, *45*, 185–213. doi:10.1006/ijhc.1996.0048

Scheiter, K., & Eitel, A. (2010). Getting a clue: Gist extraction from scenes and causal systems. In A. K. Goel, M. Jamnik, & N. H. Narayanan (Eds.), *Diagrammatic representation and inference—6th International Conference, Diagrams 2010* (LNAI 6170, pp. 264–270). Heidelberg: Springer. doi:10.1007/978-3-642-14600-8_26

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

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010b). Explaining the modality and contiguity effects: New insights from investigating students' viewing behaviour. *Applied Cognitive Psychology*, *24*, 226–237. doi:10.1002/acp.1554

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 49–69). New York: Cambridge University Press.

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

Schnotz, W., & Kürschner, C. (2008). External and internal representations in the acquisition and use of knowledge: Visualization effects on mental model construction. *Instructional Science*, *36*, 175–190. doi:10.1007/s11251-007-9029-2

Schüler, A., Scheiter, K., Rummer, R., & Gerjets, P. (2012). Explaining the modality effect in multimedia learning: Is it due to a lack of temporal contiguity with written text and pictures? *Learning and Instruction*, *22*, 92–102. doi:10.1016/j.learninstruc.2011.08.001

Schwonke, R., Berthold, K., & Renkl, A. (2009). How multiple external representations are used and how they can be made more useful. *Applied Cognitive Psychology*, *23*, 1227–1243. doi:10.1037/a0013247

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, *19*, 97–140. doi:10.1016/0364-0213(95)90005-5

Stone, D. E., & Glock, M. E. (1981). How do young adults read directions with and without pictures? *Journal of Educational Psychology*, *73*, 419–426. doi:10.1037//0022-0663.73.3.419

Ullrich, M., & Schnotz, W. (2008). Integration of picture and text: Effects of sequencing and redundancy on learning outcomes. In A. Maes, & S. Ainsworth (Eds.), *Proceedings EARLI Special Interest Group Text and Graphics: Exploiting the opportunities—Learning with textual, graphical, and multimodal representations* (pp. 148–151). Tilburg: Tilburg University.

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *The Quarterly Journal of Experimental Psychology. A*, *57*, 165–182. doi:10.1080/02724980343000189

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Van Gog, T., Kester, L., Nievelstein, F., Giesbers, B., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior*, *25*, 325–331. doi:10.1016/j.chb.2008.12.021

APPENDIX A

Text about the spatial structure and functioning of the pulley system used in the Experiment (translated from German to English):

The pulley system

The pulley system consists of three pulleys, two ropes, and one weight. The upper pulley is attached to the ceiling. Below the upper pulley is the middle pulley that is free to move up and down, and is therefore called free pulley. The upper rope is attached to the ceiling at one end, goes under the middle pulley and over the upper pulley, and is free at the other end. The lower pulley is free to move up and down, and is therefore called free pulley as well. The lower rope is attached to the ceiling at one end. It goes under the lower pulley and is attached to the middle pulley at the other end. The crate is suspended from the lower pulley. When the free end of the upper rope is pulled, the rope moves over the upper pulley and under the middle pulley, and pulls up the middle pulley. This causes the lower rope to move under the lower pulley and to pull up both the lower pulley and the crate. For these types of pulley systems, each free pulley that is added to the system splits the force with which the weight has to be lifted in half. Each free pulley that is added to the system, however, also doubles the length of rope to be pulled.