How Brief Initial Inspection of a Picture May Foster Comprehension of Text

Alexander Eitel (a.eitel@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer Str. 40,
72072 Tübingen, Germany

Katharina Scheiter (k.scheiter@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer Str. 40,
72072 Tübingen, Germany

Anne Schüler (a.schueler@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer Str. 40,
72072 Tübingen, Germany

Abstract

In the present study we hypothesized that the gist representation of a picture (extracted from brief initial inspection) supports inference generation from subsequent text, which in turn should foster comprehension. Moreover, we proposed that longer inspection of a picture is necessary to provide learners with an alternative representation that fosters mental animation and recall. Participants (*N*=76) learned from a text about pulley systems, and in three out of four conditions from an additional picture of a pulley system. Students saw either the text only, the picture preceding the text for 150ms or 2sec, or received a self-paced presentation of the picture before the text. Results confirm our assumptions that presenting the picture for the time to extract the gist (2sec) before the text fostered comprehension, whereas only the self-paced presentation of the picture fostered mental animation and recall.

Keywords: Learning from Text and Pictures; Functions of Pictures; Gist

Introduction

In research on learning from text and pictures there has been much support for the so-called multimedia principle, stating that students learn better from text and pictures than from text alone (Anglin, Vaez, & Cunningham, 2004; Mayer, 2009). According to Mayer, improved learning is reflected in improvements in retention and comprehension of the presented material. To foster comprehension, information selected from the picture has to be integrated with information selected from text (and from long-term memory) into a coherent mental representation.

An early study from Stone and Glock (1981) gives first insights about how information from text and pictures may be integrated at early stages in the learning process. In their study, text and pictures were presented simultaneously, and students initially attended to the picture for a short time prior to reading text. Stone and Glock interpreted that the brief initial glance on the picture served the purpose of extracting information about the general theme (i.e., gist).

The present study reconstructed the process of early attending to a picture for a short time before switching to reading text in an experimental design. It investigates in a systematic way whether information extracted from short initial inspection is integrated with subsequently read text, which is supposed to foster comprehension. Prior to being able to investigate which information from short inspection of the picture is integrated with text, it has to be investigated which information is actually extracted from the picture at short inspection. This has recently been done in a prior study (Eitel, Scheiter, & Schüler, 2010).

Information Extraction from Briefly Attending to Instructional Pictures

Unlike text, pictures have the property that "specific information can just be read off" (Ainsworth, 2006, p. 185), suggesting that specific information may be quickly extracted from pictures. How quickly specific information is extracted depends largely on the type of information that is extracted, as shown by a prior study (Eitel et al., 2010).

This prior study investigated early prerequisite processes in multimedia learning, namely, the question of which information is selected from (briefly attending to) instructional pictures by comparing selection processes in pictures of scenes and in instructional pictures. As instructional pictures visualizations of causal systems were used, because causal systems are commonly used as instructional material in studies on learning from text and pictures. In this prior study, students were presented with 80 pictures of scenes and 80 pictures of causal systems at four different presentation times (150ms, 600ms, 2sec, and 6sec). After presentation of each picture, students had to verify one statement about the gist of the picture, and one statement about specific details in the picture.

Results revealed that gist is rapidly extracted from pictures of scenes and causal systems. Nevertheless, it took subjects longer to achieve the same accuracy level with respect to gist extraction in causal systems (89% correct at 2sec) than in scenes (90% correct at 150ms). At these presentation times that were sufficient to extract gist in the prior study (150ms in scenes; 2sec in instructional pictures), far less information about details was extracted in both

scenes (63% at 150ms) and causal systems (63% at 2sec). Extracting information about single details (or components) thus requires longer inspection than extracting gist.

To conclude, while it takes some time to extract details (or information about single components), gist information has been shown to be rapidly extracted in pictures of scenes (at 150ms) and in instructional pictures (at 2sec). The question of whether gist information from an instructional picture already fosters learning from subsequent text is addressed in the current study by presenting the picture (for a short time) before the respective text. More specifically, students in the present study either learned from text and a picture that was presented for as long as they liked (self-paced), from text and a picture that was presented for a short time (150ms or 2sec) or from text only.

Why Learning with Multimedia May Profit from Rapid Gist Extraction

In general, students learn better from text and pictures than from text alone, because pictures in addition to text support cognitive functions that are not supported from text alone (Ainsworth, 2006; Scaife & Rogers, 1996). Pictures are especially suited to support visuo-spatial reasoning (Larkin & Simon, 1987), as it is, for instance, required when trying to understand how a causal system works (e.g., a pulley system; Hegarty, 1992).

According to Hegarty and Just (1993), understanding causal systems like pulley systems requires being able to recall its single components and being able to mentally animate the (pulley) system. Aside from recall and mental animation, we assessed comprehension of the underlying principles as a measure for comprehension on a more abstract level of organization, which we refer to as "comprehension" for reasons of simplicity in the following.

When learning about pulley systems, we assume that briefly attending to the picture is already sufficient to support some functions that learning from text alone does not support. In particular, we assume that the brief initial attention on the picture fosters comprehension, whereas no effects are expected for either mental animation or recall.

Why Briefly Attending to a Picture May Foster Comprehension

Results from the prior study suggests that watching an instructional picture (e.g., pulley system) for 2 seconds only is sufficient to get an impression about how the system roughly looks like (even though single components are not represented).

The brief initial presentation of the picture thus might free working memory resources that might otherwise be devoted to trying to visualize the (holistic) spatial structure of the system (computational offloading; Larkin & Simon, 1987). As a consequence, learners may have available more cognitive resources to engage in meaning-making activities (Ainsworth & Loizou, 2003), such as drawing inferences necessary to understand the text on a more abstract or higher level of organization (Ainsworth, 2006). Inference generation may be further supported by the fact that these inferences may be grounded in perception; that is, they are perceptually scaffolded by the (holistic) spatial structure constructed from the briefly inspected picture (Goldstone & Son, 2005; Schwartz, 1995).

Moreover, since the gist representation already gives students an impression about how the system is supposed to look like, the range of inferences that students may draw from subsequent text is already limited (graphical constraining; Scaife & Rogers, 1996). This prevents students from drawing erroneous inferences that would otherwise hinder comprehension (Schnotz & Bannert, 2003).

To conclude, we expect the gist representation of the picture (extracted after short initial presentation) to support computational offloading and graphical constraining. This should foster the generation of perceptually scaffolded inferences that, in turn, foster comprehension.

Why Briefly Attending to a Picture Does Not Foster Mental Animation

Adding a picture to a text is beneficial when the picture is better suited for problem solving than the text (re-representation; Scaife & Rogers, 1996). Mental animation mainly requires direct visuo-spatial reasoning with the picture, which is why a picture will be better suited to provide the information needed to perform mental animation than a text. Students who are presented with both a picture and a text, and are later asked to mentally animate their representation of the pulley system, will most likely profit more from information encoded in the picture than in the text. If students had enough time to encode the picture of the pulley system with respect to its single components, then students may mainly use information encoded in the picture to later perform the mental animation task (without the need to integrate information with the text).

Results from our prior study suggest that when students see a picture for 2 seconds only (or even less), they do not have enough time to encode the picture on the basis of single components (Eitel et al., 2010). Mental animation, however, requires each single component to be represented, because mental animation is a process that is carried out in a piecemeal fashion, meaning that the single components of a system are animated one after another (Hegarty, 1992, 2004). Hence, when students see the picture for 2 seconds only (or even less), they cannot solely rely on information encoded in the picture to perform mental animation in an accurate way.

To conclude, after briefly inspecting the picture, students may make use of the re-representation function (Ainsworth, 2006), which, however, will not help to better perform mental animation compared to reading text alone. Students may only benefit from re-representation once they have enough time to encode the picture on the basis of its single components.

Why Briefly Attending to a Picture Does Not Foster Recall

According to the Dual Coding Theory (Paivio, 1991), information is represented in either a verbal or a nonverbal code, whereby codes are connected to each other by referential connections enabling that information in one code can be used to retrieve information in the other code. When information is presented in two codes rather than in one, retrieval of this information is more likely because the information can be traced either by its verbal or its nonverbal code. Information encoded from a text is represented in a verbal code, whereas information extracted from a picture is represented in a nonverbal code. When a picture is added to a text, information from the text can thus be retrieved either by its verbal or – by making use of referential connections – by its nonverbal code, making correct retrieval (recall) more likely.

The model of working memory operations (Kulhavy et al., 1993) incorporates basic assumptions of Dual Coding Theory (Paivio, 1991) in that it presumes that a nonverbal code (i.e., structure of a map) can be used to activate a verbal code (i.e., facts from text). By contrast, in Kulhavy's model a special status is assigned to pictures (maps). Pictures are thereby encoded as mental images whereby their spatial structure is preserved (Kosslyn, 1995). Mental images can be processed as a single unit in working memory so that information within the image becomes simultaneously available for use. When accompanying text that describes the picture is read, facts from the text will be related to their position within the spatial structure of the mental image. Retrieval success for facts from the text is therefore increased, because to retrieve facts from the text, it is sufficient to activate the mental image of the picture. The mental image of the picture will be activated together with all its embedded information, that is, information from the picture and from the text.

To sum up, pictures foster recall from text, because they provide an additional representation that (because of its spatial structure) is especially suited to help in retrieving of information.

However, information from text and picture are connected to each other on the basis of single facts or components (Kulhavy et al., 1993; Paivio, 1991). A holistic representation (from short initial presentation) from the picture by definition does not contain single components or facts (or they are weakly represented only). Single facts in the verbal code thus cannot be related to single components from the picture. Accordingly, these facts are not better retrieved when presented with text and a picture for a short time than when presented with text only. Only when the picture can be inspected for a longer time, single facts from the text can be related to single components in the picture, and thus recall may profit from the dual coding of information.

Present Research and Hypotheses

In the present study, we investigated how comprehension, mental animation and recall are influenced by presenting pictures of pulley systems (for a short time) before the respective instructional text (see Figure 1).

By supporting computational offloading and graphical constraining, presenting the picture for 2 seconds before the text (time to extract gist) is supposed to foster comprehension. Comprehension thereby should be as good as when the presentation of the picture before the text is self-paced. Comprehension should be better when the picture is presented for 2 seconds before the text (and when presented self-paced) than when the picture is presented for 150 ms before the text and than when text only is presented (Hypothesis 1).

Since we assume that mental animation only benefits from re-representation when students inspect the picture long enough to acquire a detailed mental representation of the picture, students are assumed to profit only from the picture when they are given the time to inspect it for as long as they liked (self-paced) before the text. Students thus are assumed to neither profit from the 2-second-presentation nor from the 150ms-presentation of the picture before the text compared to the presentation of text alone. As a result, Hypothesis 2 states that mental animation is better when the presentation of the picture before the text is self-paced than when the picture is presented for a short time (2sec or 150ms) before the text and than when no picture is presented (text only). No difference between the three latter conditions is expected.

Since recall of single facts or components is assessed, and recall of single components is assumed to profit only from dual coding at longer inspection of a picture (self paced), recall should be better only when students can inspect the picture for as long as they like (before the text) compared to the presentation of text alone. The 150ms- and the 2sec-presentation of the picture before the text thus are not supposed to foster recall compared to the presentation of text only. To conclude, Hypothesis 3 states that recall is better when the presentation of the picture before the text is self-paced than when the picture is presented for a short time (2sec or 150ms) before the text and than when no picture is presented (text only). No difference between the three latter conditions is expected.

Method

Participants and Design

Seventy-six students (60 female, 16 male, average age: M = 23.93 years, SD = 3.44) from the University of Tuebingen, Germany, took part in the experiment for either payment or course credit. They were randomly assigned to one of four experimental conditions (see Figure 1): (a) text only, (b) a picture presented for 150 ms before the text (150ms-before), (c) a picture presented for 2 seconds before the text (2sec-before), or (d) a picture presented for as long as they liked

before the text (self-paced-before). There were always 19 students per condition.

Materials and Procedure

The learning material consisted of a black-and-white picture of a pulley system (Figure 1), and of a text (240 words) describing both the spatial structure of the pulley system and what happens when the rope is pulled (cf. Hegarty & Just, 1993). Moreover, two sentences were added explaining the underlying principles of pulley systems (i.e., each free pulley reduces weight to be lifted by half and doubles the length of rope to be pulled). The text contained all the information needed to understand pulley systems.

Students were tested in single sessions of approximately 30 min. They were first given a demographic questionnaire in paper-pencil format. Then students were seated in front of a computer screen. They were instructed to acquire as much information as possible from the multimedia instruction.

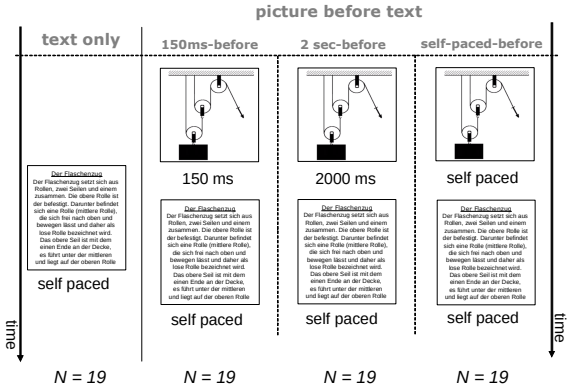

Figure 1: Design of the present experiment. Each column represents an experimental condition.

In every experimental condition, the learning phase started with a fixation cross that was displayed for 800 ms so that students could prepare for the upcoming presentation of the picture (or the text). In conditions with picture before text, the picture then appeared on the screen for either 150 ms or 2 seconds or it stayed on the screen until students signalled that they had sufficiently inspected the picture (self-paced). Then the experimenter pressed a key and the presentation of the picture was replaced by a mask that was displayed for 500 ms. Then the text appeared on the screen. In the text-only condition, the text appeared right after the fixation cross. In every experimental condition, students were first presented with text (and picture) of a toilet flush on which they were not tested, however. This "training trial" was presented so that students could familiarize themselves with the experimental procedure. The text fitted on one page, and reading was self-paced in all conditions.

Measures

We assessed comprehension of the underlying principles of pulley systems, and, based on the distinction from Hegarty (1992), we assessed mental animation and recall (of single components).

Comprehension of the underlying principles of pulley systems was assessed with both a verbal multiple choice test (3 statements; e.g., "If the weight was attached at the middle pulley, then the rope would have to be pulled with the same force than when the weight is attached to the lower pulley") and a labeling test (4 items). In the labeling test, students saw different depictions of pulley systems and they had to write down how much the to-be-lifted weight is reduced in the different pulley systems. There was always only one correct solution. Each correct solution was scored with one point. Results from both the verbal multiple choice test and the labeling test were merged in the analysis, so that students could score a maximum of seven points for comprehension of the underlying principles of pulley systems.

Mental animation and recall were assessed with a verbal multiple choice test, in which students had to verify statements about pulley systems (with yes and no). One point was given for each correct response to a statement. Nine statements tested mental animation (e.g., "if the free end of the upper rope is let go, then the middle pulley turns clockwise"), and eight statements tested recall (e.g., "both ropes are attached to the ceiling with one end"). Thus, students could score a maximum of nine points for mental animation, and a maximum of eight points for recall.

Results

To test our hypotheses, we conducted orthogonal contrast analyses following a procedure proposed by Niedenthal, Brauer, Robin, and Innes-Ker (2002), where hypotheses are translated into specific contrast. Contrasts A1 and A2 reflect the predicted pattern of results, whereas the remaining contrasts (B and C) reflect other result patterns. Accordingly, a result was considered consistent with the theoretical predictions when contrast A was statistically significant, and the remaining contrasts (B and C), as a set, were not statistically significant.

Hypothesis 1 corresponds to contrast A1 (-1 -1 1 1), meaning that we expected students in the 2sec-before and in the self-paced-before condition to show better comprehension scores than students in the text-only and in the 150ms-before condition. Given that there were four experimental conditions, two additional orthogonal contrasts captured the residual systematic variance within those conditions that were not supposed to differ among each other (B1: -1 1 0 0; C1: 0 0 -1 1). Students in the text-only condition were not supposed to differ from students in the 150ms-before condition, and students in the 2sec-before condition were not supposed to differ from students in the self-paced-before condition with regard to their comprehension scores, respectively.

Hypothesis 2 and Hypothesis 3 both correspond to contrast A2 (-1 -1 -1 3), meaning that we expected students in the self-paced-before condition to show better mental animation and recall than students in the text only, in the

150ms-before, and in 2sec-before condition. Given that there were four experimental conditions, two additional orthogonal contrasts captured the residual systematic variance within the conditions that were not supposed to differ among each other (B2: -1 -1 2 0; C2: -1 1 0 0). Students in the text-only condition were not supposed to differ from students in the 150ms-before and from students in the 2sec-before condition.

For each of the three hypotheses, all three contrasts were entered in multiple regression analyses. Contrast A was entered as a single predictor, and the two remaining contrasts (B and C) were entered as a set in the multiple regression analysis.

Hypothesis 1 was supported (see Figure 2). A multiple regression analysis in which comprehension of the underlying principles of pulley systems was regressed on all three contrasts revealed that contrast A1 was statistically significant, $F(1, 72) = 12.17$, $p = .001$. As expected, students in the 2sec-before and in the self-paced-before condition had better comprehension scores than students in the 150ms-before and in the text-only condition. The two remaining contrasts (B1 and C1), as a set, were not significant, $F < 1$, meaning that comprehension scores did neither differ between students in the 2sec-before and the self-paced-before condition nor between students in 150ms-before and the text only condition.

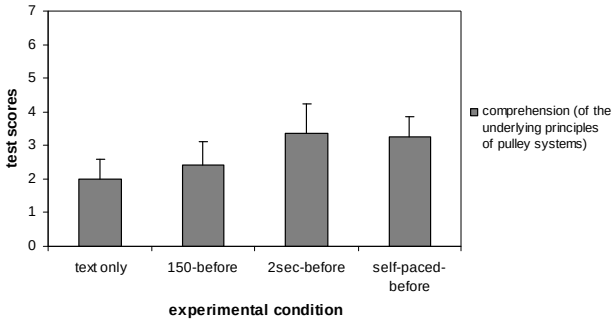

Figure 2: Results for comprehension of the underlying principles of pulley systems.

Hypothesis 2 was supported (see Figure 3). A multiple regression analysis in which mental animation performance was regressed on all three contrasts revealed that contrast A was statistically significant, $F(1, 72) = 20.78$, $p < .001$. As expected, mental animation was better in the self-paced-before than in the 2sec-before, than in the 150ms-before and than in the text-only condition. The latter three conditions did not significantly differ among each other, $F(1, 72) = 2.44$, $p = .12$. Hypothesis 3 was supported (see Figure 3). A multiple regression analysis in which recall performance was regressed on all three contrasts revealed that contrast A was statistically significant, $F(1, 72) = 5.36$, $p = .02$. As expected, recall was better in the self-paced-before than in the 2sec-before, than in the 150ms-before and than in the text-only condition. The other three conditions did not significantly differ among each other, $F < 1$.

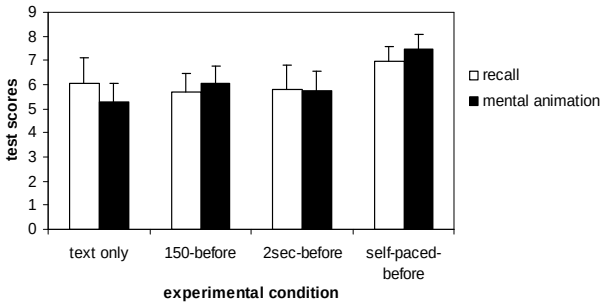

Figure 3: Results for recall and mental animation performance.

To sum up, comprehension (of the underlying principles of pulley systems) was better in the 2sec-before and in the self-paced-before than in the 150ms-before and in the text-only condition, yielding support for Hypothesis 1. Recall and mental animation were better in the self-paced-before compared to the all three other conditions, supporting Hypothesis 2 and Hypothesis 3.

Students in the self-paced before condition inspected the picture longer (than in the 2sec-before condition. Moreover, reading time was shorter in the self-paced before than in the text only condition, but did not differ between the remaining conditions. Reading time was not correlated to performance.

Summary and Discussion

In the present study we first hypothesized that briefly attending to a picture fosters comprehension by rapidly providing holistic information (Eitel et al., 2010) that is sufficient to support inference generation through computational offloading and graphical constraining (Scaife & Rogers, 1996). Results confirmed Hypothesis 1, because students had better comprehension (i.e., of the underlying principles of pulley systems) when they saw a picture of a pulley system for 2 seconds before they learned from the respective text than when they received text only or the picture for 150 ms before the text.

According to Hypothesis 2, mental animation only profits from re-representation when the picture is inspected long enough so that single components are encoded and represented in memory, making the piecemeal mental animation process possible (Hegarty, 1992, 2004). Only the self-paced presentation of the picture led to better mental animation, confirming Hypothesis 2.

Since only longer inspection of the picture was supposed to lead to dual coding of the single components, which is necessary to foster recall (of single components), Hypothesis 3 stated that recall should be better in the self-paced before compared to the other three conditions. This was confirmed by the results as well.

To conclude, by considering the nature of the functions of pictures (Ainsworth, 2006; Scaife & Rogers, 1996), the present study showed that specific predictions can be made about when a picture (e.g., presented for 2sec before text) fosters comprehension in addition to learning from text.

Implications and Further Research

A short initial glance at a picture can already foster comprehension of subsequent text! Therefore it may be an effective strategy to first attend to the picture for a short time before starting to read the text. Students actually showed this type of processing in an early study of Stone and Glock (1981). While reading text, students switched back and forth between text and picture quite often. According to Hegarty and Just (1993), these switches are important to integrate information from text and picture on the basis of single components (or units of components) so that learners are able build a comprehensive mental model of the multimedia message, which in turn fosters comprehension. To what degree briefly attending to the picture fosters comprehension compared to later switches to the picture while reading is not yet clear. Further studies will be needed to assess the relative importance of the brief initial inspection of the picture.

In the present study we assessed effects of the short presentation of the picture before the text with one type of instructional material (i.e., pulley systems). Instructional pictures of pulley systems in the present study were simple line drawings in black-and-white. Further studies with different instructional material and more complex instructional pictures will be needed to see whether the beneficial effect of the short initial presentation of the picture can be generalized. One would assume that the short initial presentation of the picture only fosters comprehension when the spatial structure is rather easy to encode (so that it is rapidly extracted), and when the spatial structure is relevant to the functioning of the system.

Finally, there is reason to assume that the gist representation of the picture contains the spatial structure of the picture (spatial scaffold; Castelhano & Henderson, 2007), which likely has been responsible for the better comprehension of subsequent text in the present study. If a spatial scaffold indeed is sufficient to foster comprehension of subsequent text, then the presentation of an externally presented scaffold (i.e., perceptually degraded or incomplete picture of a pulley system) should have the same effect. We are currently conducting further studies that check for this assumption.

References

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

Ainsworth, S. E., & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27*, 669-681.

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.

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.

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.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences, 14*, 69–110.

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

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in cognitive science, 8*, 280-285.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*, 717-742.

Kosslyn, S. M. (1995). Mental imagery. In S. M. Kosslyn & D. N. Osherson (Eds.), *An invitation to Cognitive Science: Visual cognition* (Vol 2, 2nd ed., pp. 267-296). Cambridge, MA: MIT Press.

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.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65-99.

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

Niedenthal, P. M., Brauer, M., Robin, L., & Innes-Ker. Å. H. (2002). Adult attachment and the perception of facial expression of emotion. *Journal of Personality and Social Psychology, 82*, 419-433.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology, 45*, 255–287.

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

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representations. *Learning and Instruction, 13*, 141-156.

Schwartz, D. L. (1995). Reasoning about the referent of a picture versus reasoning about the picture as the referent: An effect of visual realism. *Memory & Cognition, 23*, 709–722.

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