Processing multimedia material: Does integration of text and pictures result in a single or two interconnected mental representations?

Anne Schüler*, Jana Arndt, Katharina Scheiter
Knowledge Media Research Center, Tuebingen, Germany

A R T I C L E   I N F O
Article history:
Received 20 December 2012
Received in revised form 19 September 2014
Accepted 25 September 2014
Available online 16 October 2014

Keywords:
Multimedia learning
Integration of text and pictures
Misinformation effect
Mental representations
Dual coding

A B S T R A C T
We investigated whether individuals construct either two interconnected mental representations or a single mental representation that merges information from two representations when processing multimedia materials. Individuals memorized text-picture stimuli in four different versions that differed in the specificity of information contained in either text or pictures: general pictures/general sentences, general pictures/specific sentences, specific pictures/general sentences, and specific pictures/specific sentences. Afterwards, individuals decided whether they had previously seen the specific or the general version of the sentences and the pictures. Across two experiments, individuals more frequently falsely recognized the specific sentences after having seen general sentences/specific pictures. This indicates that individuals had integrated the specific picture information with the general sentence information into a single mental representation. No such effects were observed for picture recognition. The implications of these results for multimedia learning are discussed.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

In the past decades, it has been demonstrated repeatedly that people learn more deeply from text and pictures than from text alone (for overviews see Levie & Lentz, 1982; Mayer, 2009). Multimedia learning theories explain this finding by assuming that text and pictures are integrated with each other (e.g., Mayer, 2009; Schnitz & Bannert, 2003). The resulting mental representation is assumed to better enable learners to recall and apply the learnt content to novel situations. However, despite its importance for multimedia learning (e.g., Bodemer, Ploetzner, Feuerlein, & Spada, 2004; Seufert, 2003), the precise nature of the integration process has never been examined empirically under conditions of multimedia learning. This was the aim of the studies reported here. From a theoretical perspective, knowing more about the kind of mental representations constructed during processing multimedia materials is pivotal to understand how people process text and pictures. In addition, as Renkl (2013) points out, “educational research is typically not only theoretically but also practically grounded” (p. 318). Thus, knowing more about the way people learn from text and pictures allows us to derive recommendations for the design of multimedia materials that are grounded in cognitive theory (e.g., Mayer, 2009; for an analogous example see also Rummer, Schwepp, Fürstenberg, Scheiter, & Zindler (2011) and related follow-up research by Leahy & Sweller (2011)).

In the reported studies we contrasted two views concerning the integration of text and pictures. According to the first view two separate mental representations are constructed (i.e., one from the text and one from the picture), which are interconnected (i.e., integrated) via referential connections. We will call this view the “two-mental-representations” view, which goes back to dual coding theory (Paivio, 1990). According to the second view a single mental representation is constructed in which the information from the two representations is merged (i.e., integrated). We will call this view the “single-mental-representation” view, which goes back to theories on mental model construction (Johnson-Laird, 1983; Kintsch, 1988).

In the following, we will first introduce the assumptions regarding the integration process made by the two most prominent theories of multimedia learning, namely, the cognitive theory of multimedia learning (Mayer, 2009) and the integrated model of text and picture comprehension (Schnitz & Bannert, 2003), before describing the aforementioned views in more detail.

* Corresponding author. Knowledge Media Research Center, Schleichstraße 6, 72076 Tuebingen, Germany. Tel.: +49 7071 979341.
E-mail address: a.schueler@wmm-kmrc.de (A. Schüler).
1.1. Theoretical assumptions of multimedia theories regarding text-picture integration

The cognitive theory of multimedia learning (Mayer, 2009) posits that when text and pictures are presented together, learners construct a verbal representation from the text as well as a pictorial representation from the picture, which are then integrated. Concerning the integration process, Mayer (1997) assumes that learners build one-to-one correspondences (i.e., referential connections) between the verbal and the pictorial mental representation. This assumption is theoretically grounded in dual coding theory (Paivio, 1990) and appears to be in line with the two-mental-representations view. On the other hand, Mayer (2009, p. 77) states that making connections between the verbal and pictorial mental model “involves a change from having two separate representations — a pictorial model and a verbal model — to having an integrated representation.” Furthermore, in Mayer’s diagrammatic depiction of the cognitive theory of multimedia learning, the resulting mental model is represented by means of a single element in working memory (see for example, Mayer, 2009, p. 77). This assumption seems to better correspond with the single-mental-representation view.

The integrated model of text and picture comprehension (Schnotz & Bannert, 2003) assumes that individuals construct a propositional representation from the text and a mental model from the picture. The propositional representation may trigger the construction of the mental model; furthermore, the mental model can be used to read off information in a propositional format. As the authors point out, constructing a propositional representation as well as a mental model seems to be, at first glance, similar to dual coding (Schnozt & Bannert, 2003). However, in contrast to dual coding theory, the authors do not assume that the mental model simply adds an additional pictorial code that yields a quantitative advantage compared to a single code. Instead, the mental model is a qualitatively different representation, which incorporates also information from text. Thus, the assumptions of Schnozt and Bannert (2003) are more in line with the single-mental-representation view than with the two-mental-representations view.

To conclude, both theories agree that the integration of text and pictures is crucial to learning from multimedia. However, whether the integration process results in two mental representations, which are interconnected with each other (Mayer, 1997), or a single representation, which merges information from text and picture (Mayer, 2009; Schnozt & Bannert, 2003), is still an open question. Therefore, the aim of the studies reported in the current paper was to examine the precise nature of this integration process under conditions of multimedia learning. In the following, we will describe the two different views concerning the integration of text and pictures in more detail.

1.2. Construction of mental representations when processing text and pictures

According to the two-mental-representations view, text and pictures support the construction of two mode-specific mental representations, which are interconnected via referential connections. This view is grounded in dual coding theory (Paivio, 1990). According to this theory, the cognitive system is composed of a verbal system, which deals with language, and a nonverbal system, which deals with the analysis of pictorial information and the generation of mental images. One system can trigger activity in the other system by means of referential connections between representations stored in them. For example, when a picture is presented, the corresponding verbal representation can be activated via activation spreading. Dual coding theory assumes that dual coding of information yields better recognition and recall performance because if one code is forgotten the other code may still be accessible in memory (e.g., Paivio, 1965).

According to the two-mental-representations view based on dual coding theory, text-picture presentations are more likely to lead to two mental representations interconnected via referential connections than presenting either text or pictures, which can explain why people learn more deeply from multimedia materials.

On the other hand, according to the single-mental-representation view, text and pictures support the construction of a single mental representation that contains verbal as well as pictorial information. This view is grounded in theories proposing that during information processing, individuals construct comprehensive mental representations known as situation models (Kintsch, 1988) or mental models (Johnson-Laird, 1983; Zwaan & Radvansky, 1998). According to these theories a single mental representation is constructed, which merges information derived from different information sources and prior knowledge (for empirical evidence, see for example Allen, Hitch, & Baddeley, 2009; Elsley & Parmentier, 2009; Maybery et al., 2009). It is assumed that this mental representation comprises all the information from the different external representations (e.g., Baddeley, Allen, & Hitch, 2011; Brunyé & Taylor, 2008; Payne & Baguley, 2006; Zwaan & Radvansky, 1998). This implies that by integrating information across non-redundant information sources, the resulting mental representation becomes more specific. For example, when a picture of a car is presented and a sentence additionally states that the car has ski-racks on it, it is assumed that the mental representation will then contain information about a car with ski-racks (see Pezdek, 1977). Conversely, it is not assumed that the information from one external representation overwrites the information conveyed by another external representation. Thus, the mental representation will not only contain the picture information (i.e., a car without ski-racks).

To conclude, regarding text-picture processing, mental model research suggests that a single mental representation is constructed that contains information from the text as well as from the picture. When either text or pictures are presented, the mental representation contains only the information from one source.

1.3. Empirical evidence for the two-mental-representations and single-mental-representation view

So far, there are no studies to our knowledge that have attempted to answer the question of which kind of mental representation is constructed when text and pictures are integrated with each other under multimedia learning conditions; moreover, there are very few studies that investigated text-picture integration in other contexts. Of course, many studies exist in which text and pictures were presented to individuals and performance measures were applied afterwards (e.g., Bransford & Johnson, 1972; Mayer & Anderson, 1991; for an overview, see Mayer, 2009). In these studies, performance has often been interpreted as evidence for the existence of an integrated representation. However, the exact nature of the underlying mental representation has not been explored any further. Other studies have tried to investigate the exact nature of the mental representations more explicitly; however, although the results indicated that text and pictures were integrated with each other, they did not aim at testing and were thus inconclusive regarding the two-mental-representations view versus single-mental-representation view (e.g., Bower & Morrow, 1990;
Glenberg & Langston, 1992; McNamara, Halpin, & Hardy, 1992; Morrow, Bower, & Greenspan, 1989). More conclusive evidence for the kind of mental representation constructed during text-picture processing comes from early research concerning the misinformation effect (e.g., Gentner & Loftus, 1979; Pezdek, 1977; Wippich, 1987). In the studies reported below, we adapted the paradigm used in that early research to conditions of multimedia learning.

Gentner and Loftus (1979) presented their individuals with pictures that were either general or specific. For example, a general picture showed a girl walking down a path. The corresponding specific picture was identical to the general picture, except that the girl was hiking instead of walking; that is, the girl in the specific picture had a backpack, boots, and a hiking stick. Thus, the specific information always contained the entire representation of the general information (e.g., walking), but also comprised additional information (e.g., hiking). Each individual had a sheet of paper containing a list of 39 sentences. The individuals’ task was to match each picture to one of the sentences. The list always contained only one sentence that matched the respective picture. In addition to manipulating the picture’s specificity, the matching sentences’ specificity was also varied. That is, the sentences were either general (e.g., “The girl is walking down the path”) or specific (e.g., “The girl is hiking down the path”). Importantly, individuals were presented with only one of the four possible picture-sentence combinations for each item (i.e., general picture/general sentence; general picture/specific sentence; specific picture/general sentence; specific picture/specific sentence). After one week, individuals solved a forced-choice picture recognition test. In this test they were presented with the general and specific pictures and had to decide which member of each pair they had seen the week before. Gentner and Loftus assumed that “specific information presented in a sentence would be added to the subject’s representation of the picture to produce a more specific picture representation” (p. 366). In line with this assumption, Gentner and Loftus showed that individuals who had received general pictures together with specific sentences had a higher probability of falsely recognizing the specific pictures than individuals who had received general pictures together with general sentences. These findings can be explained by the fact that the specific sentence information (e.g., hiking) had been integrated with the general picture (e.g., a girl walking down a path), leading students to falsely recognize the specific picture in the post-test. Thus, the data support the assumptions of the single-mental-representation view. The results of Gentner and Loftus were replicated by Wippich (1987) for picture as well as sentence recognition. Finally, Pezdek (1977) obtained similar results by inserting specific objects into the general information (e.g., a general picture of a car without any accessories next to a tree and a specific sentence stating that “the car by the tree had ski-racks on it”) instead of specifying verbs (e.g., Gentner & Loftus, 1979).

In sum, the results by Gentner and Loftus (1979), Wippich (1987), and Pezdek (1977) speak in favor of the assumption that individuals integrate pictorial and verbal information with each other into one coherent mental representation, thereby supporting the single-mental-representation view.

Importantly, the two-mental-representations view cannot account for these results. For example, if the verbal representation contained specific information about ski-racks but the pictorial information did not contain such information, the probability that the verbal information would be connected via referential connections to a pictorial, nonverbal representation of ski-racks would be low. This should only be the case if individuals construct a second (pictorial) code on their own when, for instance, reading about a car with ski-racks on it. In other words, according to the two-mental-representations view specific verbal information about ski-racks is less likely to be dual coded (i.e., low probability that a specific pictorial representation containing ski-rackets is constructed). Thus, individuals should not falsely recognize the specific picture when originally a general picture had been combined with a specific sentence. This, however, was demonstrated by the aforementioned studies.

Despite the fact that the results observed within the context of research on the misinformation effect speak in favor of the single-mental-representation view, they cannot be unambiguously transferred to multimedia learning conditions. This is due to two reasons: First, the studies suffer from some methodological weaknesses (e.g., only picture but not sentence recognition was measured (Gentner & Loftus, 1979); use of verification items that might prevent participants from noticing possible deviations from the original item (Pezdek, 1977); pictures were presented only for short time intervals (e.g., 2 s; Gentner & Loftus, 1979); recognition was measured after a long test delay only (Gentner & Loftus, 1979; Wippich, 1987)). Second, these studies were conducted from a different perspective; the aim of these studies was not to specify the nature of the integration process under conditions of multimedia learning. Consequently, none of these studies formulated the hypotheses based on the central assumptions in multimedia learning (i.e., single-mental representation vs. two-mental-representation) and none of the studies incorporated all of the relevant multimedia learning conditions. Especially the fact that participants in these studies had a high affordance to process the materials (e.g., by searching actively for corresponding sentences on a sheet of paper while pictures were presented on slides, see Gentner & Loftus, 1979; Wippich, 1987) sequential text-picture presentation, see Pezdek, 1977) makes it difficult to predict what will happen under multimedia learning conditions, where the pace of presentation is most often controlled by the system (e.g., Brünken, Steinbacher, Plass, & Leutner, 2002; Mayer & Anderson, 1991) and pictures and sentences are normally presented simultaneously on the computer screen. This simultaneous presentation of information is assumed to support the integration process because making connections between text and pictures (two-mental-representations view) or merging them into a single mental representation (single-mental-representation view) should be facilitated if both representations are available at the same time (e.g., Ginn, 2006; Mayer, 2009). Accordingly, the evidence obtained from studies conducted in the context of the misinformation effect cannot be transferred unambiguously to multimedia learning conditions. Therefore, the aim of the conducted studies was to disentangle the single-mental-representation view from the two-mental representation view under conditions of multimedia learning.

1.4. The present study: overview and hypotheses

In two experiments, individuals were first asked to memorize simultaneously presented text-picture stimuli in four different versions: general pictures/general sentences, general pictures/specific sentences, specific pictures/general sentences, and specific pictures/specific sentences. Importantly, the specific information (e.g. car with ski-racks) always contained the entire representation of the general information (e.g., car), but added additional information (e.g., ski-racks). After the presentation phase, general and specific pictures as well as general and
specific sentences were presented in a forced-choice format and individuals decided which one of the two items they had seen previously. In line with Gentner and Loftus (1979), we used the frequency of choosing a specific picture (or specific sentence) as the dependent measure.

The advantage of this operationalization was that we were able to test precise predictions regarding the specific patterns of results expected from the two-mental-representations view versus single-mental-representation view.

Both views predict that individuals will correctly decide against the specific item in the post-test (picture or sentence) after having encoded general pictures/general sentences. Furthermore, they will correctly decide in favor of the specific item in the post-test after having seen specific pictures/specific sentences. This is because according to the two-mental-representations view, in both cases the conveyed information (general or specific) is dual coded and therefore accessible in memory. According to the single-mental-representation view, in both cases the mental model merges only general or only specific information from text and picture (see Fig. 1).

The critical comparisons for which the single-mental-representation view and the two-mental-representation view make opposing predictions involve the two mismatch conditions of general pictures/specific sentences and specific pictures/general sentences.

Regarding picture recognition, the single-mental-representation view predicts that individuals will more often falsely decide in favor of the specific picture after having seen general pictures/specific sentences than after having seen general pictures/general sentences (hypothesis 1.1). This is because the mental representation merges general picture and specific sentence information, leading to a higher frequency of false

<table>
<thead>
<tr>
<th>Picture</th>
<th>Sentence</th>
<th>Two-mental-representations view</th>
<th>Single-mental-representation view</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>General</td>
<td>Picture information</td>
<td>Sentence information</td>
</tr>
<tr>
<td><img src="Image1" alt="Table 1" /></td>
<td><img src="Image2" alt="Table 2" /></td>
<td>dual coded</td>
<td>dual coded</td>
</tr>
<tr>
<td>Specific</td>
<td>Specific</td>
<td>Der Schreibtisch steht unter dem Fenster</td>
<td>dual coded</td>
</tr>
<tr>
<td><img src="Image3" alt="Table 3" /></td>
<td><img src="Image4" alt="Table 4" /></td>
<td>overwritten</td>
<td>available</td>
</tr>
<tr>
<td>General</td>
<td>Specific</td>
<td>Der Schreibtisch steht unter dem Fenster</td>
<td>dual coded</td>
</tr>
<tr>
<td><img src="Image5" alt="Table 5" /></td>
<td><img src="Image6" alt="Table 6" /></td>
<td>Falsely deciding against specific sentence</td>
<td>available</td>
</tr>
<tr>
<td>Specific</td>
<td>General</td>
<td>Der Tisch steht unter dem Fenster</td>
<td>single coded</td>
</tr>
<tr>
<td><img src="Image7" alt="Table 7" /></td>
<td><img src="Image8" alt="Table 8" /></td>
<td>Falsely deciding against specific sentence</td>
<td>available</td>
</tr>
</tbody>
</table>

Fig. 1. Text-picture combinations and predictions of the two-mental-representations view versus single-mental-representation view.
decisions in favor of the specific picture, whereas in the latter case the mental representation contains only the general information from both information sources, leading to fewer false decisions. On the contrary, the two-mental-representations view leads to a very different prediction. It assumes that individuals will more often falsely decide against the specific picture after having seen specific pictures/general sentences than after having seen general pictures/specific sentences (hypothesis 1.2). This is because in the former case the specific information is not dual coded and therefore less accessible in memory, whereas in the latter case the specific information from the picture is also included in the sentence and therefore dual coded (see also Fig. 1).

The same rationale holds for sentence recognition: Here, the single-mental-representation view predicts that individuals will more often falsely decide in favor of the specific sentence after having seen specific pictures/general sentences than after having seen general pictures/general sentences (hypothesis 2.1). On the contrary, the two-mental-representations view does not predict such differences but presumes that individuals will more often falsely decide against the specific sentence after having seen general pictures/specific sentences than after having seen specific pictures/specific sentences (hypothesis 2.2; see also Fig. 1).

Next to these main hypotheses, based on the data reported by Gentner and Loftus (1979), a main effect of specificity was expected; that is, individuals should choose the specific picture (or sentence) more often when a specific picture (or sentence) was presented than when a general picture (or general sentence) was presented. Because we used contrast analyses to test our hypotheses, it was necessary to specify the whole expected data pattern, including this main effect.

2. Experiment 1

Prior to Experiment 1, we conducted a pre-study with 44 participants (37 female; average age: M = 21.93 years, SD = 3.14) to test whether the sentence-picture correspondence varied between the four possible sentence-picture combinations for each item. This was necessary to exclude the possibility that in the later experiments differences between conditions were due to differences in the perceived match of sentence-picture combinations. For each of 42 sentence-picture stimuli a general versus specific sentence as well as a general versus specific picture were generated, yielding 4 possible sentence-picture combinations for each item. Four different versions of a questionnaire were developed, with each version containing only one type of sentence-picture combination for each of the 42 items. Eleven subjects answered each questionnaire version. The subjects were instructed as follows: “Please rate on a 5-point scale each of the following sentence-picture combinations in terms of how well the sentences and pictures match each other.” The scale ranged from (1) not at all to (5) very well. For each item, ANOVAs were calculated with sentence-picture combination as the between-subjects factor and perceived match as the dependent variable. The analyses revealed significant differences in the perceived match for 10 items, which were excluded from the item pool as a consequence. Thus, for Experiment 1, a pool of 32 sentence-picture combinations was available. The ratings of these items ranged in all four conditions from 1 to 5. However, importantly, the ratings of the single items did not differ between conditions, indicating that if subjects rated sentence-picture combinations as less matching, this was due to factors other than the information that differed between items.

2.1. Method

2.1.1. Participants and design

Forty-seven participants (31 female, average age: M = 22.74 years, SD = 3.68) with varied educational backgrounds participated in the study for either payment or course credit. The within-subjects factor sentence-picture combinations contained four levels: general sentences/general pictures, general sentences/specific pictures, specific sentences/general pictures and specific sentences/specific pictures.

2.1.2. Material

The items consisted of 32 sentence-picture combinations; that is, eight sentence-picture combinations per condition. The pictures were simple black-and-white line drawings and depicted different scenes or objects (e.g., a boat on the sea, a teacher in front of a blackboard, a pie on a table). The pictures referred to the pictures and were four to nine words long. They were presented below the pictures. Depending on condition, the pictures depicted either general or more specific information. For example, a general picture showed a table below a window, whereas the corresponding specific picture showed a desk below the window, that is, a table with pencils and a desk lamp on it (see Fig. 1; see Appendix for more examples). Specific pictures were generated by adding details to the general picture.

The general and more specific sentences were generated analogously. These sentence versions were always identical except that the specific sentence contained a noun that was more specific than the noun used in the general sentence (see Psyzek, 1977). Specific and general nouns described objects of the same category; however, the specific noun always referred to a more specific object of that category. For example, a general sentence stated that “the table is below the window”, whereas the corresponding specific sentence stated that “the desk is below the window” (see Appendix for more examples). The nouns were chosen from a German word list — called “DeReWo-Grund-/Wortformenliste” — that lists each word’s relative frequency. We defined general nouns as more frequently used German words compared to specific nouns because it can be assumed that the degree of specificity corresponds with frequency of use.

2.1.3. Measures

The measures included a picture and a sentence recognition forced-choice test. The picture recognition forced-choice test contained 16 out of the 32 pictures presented in the presentation phase. These pictures were selected randomly from the item pool and were the same for each participant. For each picture, the general and the specific picture version were presented next to each other on one slide. The order of the general and specific pictures was randomized. Subjects’ task was to decide which one of the pictures they had seen earlier. In line with Gentner and Loftus (1979), participants received one point for choosing the specific picture, resulting in a maximum score of 16 points for the picture recognition forced-choice test. Importantly, this score reflects false alarms or hits depending on the sentence-picture combination shown in the presentation phase. Thus, when subjects had been presented with general pictures (paired with either specific or general sentences) in the presentation phase, choosing a specific picture was erroneous (i.e., a false alarm), whereas when subjects had been presented with specific pictures (paired with either specific or general sentences) in the presentation phase, choosing a specific picture was correct (i.e., a hit).

The sentence recognition forced-choice test was constructed in the same manner as the picture recognition forced-choice test: It
contained 16 out of the 32 sentences presented in the presentation phase. These sentences were part of the items that remained in the item pool after randomly selecting the items for the picture recognition test: The sentences whose corresponding picture had been used in the picture recognition task were not used in the sentence recognition task and vice versa. For each sentence, the general and the specific sentence version were presented next to each other. The order of general and specific sentences was randomized. Subjects had to decide which of the two sentences they had seen during the presentation of the material. Participants received one point for choosing the specific sentence, resulting in a maximum score of 16 points for the sentence recognition forced-choice test.

The scores for specific picture and specific sentence selection were converted into percentages for easier interpretation.

2.1.4. Procedure

Participants were tested in groups with up to five participants. Each participant completed the experiment individually. They were seated in front of a laptop with a display resolution of 1280 × 800 pixels. First, they were given a short computer-based written instruction concerning the duration and the procedure of the experiment. Second, they were told that 32 sentence-picture combinations would be presented, that the presentation was system-paced, and that their task was to attend carefully to the stimuli and to memorize each sentence-picture combination the best they could. The sentence-picture combinations were then presented randomly on a 15-inch screen with the software E-Prime 2.0 Professional from Psychology Software Tools®. Each sentence-picture combination was presented individually and remained visible on the computer screen for 8 s. This time interval was a compromise between the short presentation times used in research on the misinformation effect (ranging from two to 7 s, see Gentner & Loftus (1979) and Pezdek (1980)) and the longer presentation times typically deployed in multimedia studies. After that, a white screen appeared for 2 s before the next sentence-picture combination was presented. The presentation phase took about 7 min. After the presentation phase, participants took a short break of about 15 min during which they worked on two tasks measuring their verbal and spatial working memory capacity, respectively. Because there were no relations between these measures and recognition performance, we refrain from describing them in more detail. The participants then answered the sentence recognition test followed by the picture recognition test. No time restrictions were given for the post-tests. Finally, they answered demographic questions about their age, gender, and field of study. A single experimental session took about 30–45 min.

2.2. Results

Table 1 shows the relative frequency of choosing the specific picture and the specific sentence in each of the four within-subjects conditions. Thus, for example, the third row of Table 1 indicates that when individuals had been presented with general pictures and general sentences, they chose the specific sentence in 9.04% of the cases, whereas when individuals had been presented with specific pictures and general sentences, they chose the specific sentence in 14.89% of the cases.

To test the two-mental-representations view versus the single-mental-representation view, we first conducted one-factor repeated measures ANOVAs with sentence-picture combination (general picture/general sentence; general picture/specific sentence; specific picture/general sentence; specific picture/specific sentence) as a within-subjects independent variable. Significant effects of the within-subject factor sentence-picture combinations were then followed up by planned contrasts to test the patterns of results (see Field, 2005). Note that the four conditions would also allow for calculating 2 × 2 within-subjects ANOVAs. However, because both views predict specific interactions, which cannot be tested by unspecific ANOVAs, contrast analyses are much better suited (see Furr & Rosenthal, 2003; Nückles, Hübner, & Renkl, 2009).

With regard to the picture recognition forced-choice test, the within-subjects ANOVA revealed a significant effect for sentence-picture combinations, F(3, 138) = 193.96, MSE = 193.96, p < .001, η2 = .81. To test whether the differences between sentence-picture combinations were in line with either the two-mental-representations view or single-mental-representation view, three planned contrasts were conducted.

The first planned contrast revealed no difference between general pictures/general sentences and general pictures/specific sentences, F(1, 46) = 1.14, MSE = 570.07, p = .29, η2 = .02, contradicting hypothesis 1.1, that is, the assumption of the single-mental-representation view. The second planned contrast revealed no difference between the presentations of specific pictures paired with specific versus general sentences, F < 1, contradicting hypothesis 1.2, that is, the assumption of the two-mental-representations view. The third planned contrast revealed that individuals chose a specific picture more often after previously having seen the specific picture than after having seen a general picture, F(1, 46) = 410.10, MSE = 7839.21, p < .001, η2 = .90 (M = 87.50%, SE = 1.93 vs. M = 22.07%, SE = 3.1), in line with the assumed main effect. In sum, the pattern of results for the picture recognition test did not allow disentangling the two-mental-representations view from the single-mental-representation view.

With regard to the sentence recognition forced-choice test, the repeated ANOVA revealed a significant effect for the within-subjects factor sentence-picture combinations, F(3, 138) = 172.60, MSE = 377.88, p < .001, η2 = .79. This time, the first planned contrast revealed that individuals more often falsely chose the specific sentence after having seen general sentences/specific pictures than after having seen general sentences/general pictures, F(1, 46) = 4.84, MSE = 331.87, p = .03, η2 = .10, supporting hypothesis 2.1 (i.e., the single-mental-representation view). The second planned contrast revealed no difference between specific sentences/general pictures and specific sentences/specific pictures, F < 1, contradicting hypothesis 2.2 (i.e., the two-mental-representations view) and therefore supporting again the single-mental-representation view, which predicted no differences. The third planned contrast revealed that individuals chose a specific sentence more often after having seen the specific sentence than after having seen the general sentence in the presentation phase (M = 76.33%, SE = 2.64 vs. M = 11.97%, SE = 2.15; F(1, 46) = 318.51, MSE = 9789.30, p < .001, η2 = .87).
2.3. Summary and discussion of Experiment 1

To summarize, the results of Experiment 1 confirmed the hypothesized pattern of results of the single-mental-representation view for the sentence recognition forced-choice test indicating that specific picture information was integrated with general sentence information, leading participants to falsely choose specific sentences. Regarding the picture recognition forced-choice test, neither view was confirmed. One explanation for the latter result might be the high performance levels observed for some of the items in the pictorial forced-choice test (i.e., 25% of the items were recognized in more than 90% of the cases). However, because the item pool of Experiment 1 was rather small, it was not possible to exclude these items from analyses. Therefore, we decided to develop more items for Experiment 2 so that it would be possible to exclude items from analyses in case they were too easy and, therefore, inappropriate for revealing differences between sentence-picture combinations.

3. Experiment 2

We conducted a pre-study with 29 participants (21 female, average age: $M = 24.07$ years, $SD = 3.33$) to test whether the correspondence of the 48 newly created sentence-picture stimuli varied between the four possible sentence-picture combinations for each item. Four different versions of a questionnaire were developed, with each version containing only one type of sentence-picture combination for each of the 42 items. Seven to 8 subjects completed each version of the questionnaire, which asked them to rate on a 5-point rating scale how well each sentence and picture matched. ANOVAs with sentence-picture combination as a between-subjects factor and perceived match as the dependent variable revealed (marginally) significant differences in the perceived match for 15 items that were then excluded from the item pool. Thus, for Experiment 2, 33 sentence-picture combinations were available in addition to those from Experiment 1.

3.1. Method

3.1.1. Participants and design

Seventy-five participants (52 female, average age: $M = 24.23$ years, $SD = 3.02$) with different educational backgrounds participated in the study for either payment or course credits. The within-subjects factor sentence-picture combinations contained four different types of sentence-picture combinations: general sentences/general pictures, general sentences/specific pictures, specific sentences/general pictures, and specific sentences/specific pictures.

3.1.2. Material

Fifty-nine sentence-picture combinations (33 new items, 26 items from Experiment 1) were presented. There were either 14 or 15 sentence-picture combinations per condition.

3.1.3. Measures

The forced-choice recognition tests were identical to those in Experiment 1 except for the number of items: The picture recognition forced-choice test contained 29 picture pairs; the sentence recognition forced-choice test contained 30 sentence pairs. Again, participants received one point for choosing the specific stimulus version; points were converted into percentages.

3.1.4. Procedure

The procedure was the same as in the first experiment, except for the tasks performed in the 15 min break. This time, participants filled in questionnaires which were related to another experiment conducted subsequently, but were unrelated to the present study.

3.2. Results

Fourteen items had to be excluded from analyses due to ceiling effects (that is, a recognition rate higher than 90%). The remaining 45 items were analyzed by one-factor repeated measures ANOVAs with sentence-picture combinations as the within-subjects independent variable, followed-up by the planned contrasts. The dependent variable was the relative frequency of choosing the specific picture or specific sentence. The descriptive data are shown in Table 2.

With regard to the picture recognition forced-choice test, the within-subjects ANOVA revealed a significant effect for sentence-picture combinations, $F(3, 222) = 340.75, MSE = 293.35, p < .001, \eta_p^2 = .82$. The first planned contrast revealed no difference between general pictures/general sentences and general pictures/specific sentences, $F < 1$, contradicting hypothesis 1.1, that is, the assumption of the single-mental-representation view. The second planned contrast revealed no difference between the presentations of specific pictures paired with specific or general sentences, $F < 1$, contradicting hypothesis 1.2, that is, the assumption of the two-mental-representations view.

With regard to the sentence recognition forced-choice test, the repeated ANOVA revealed a significant effect for sentence-picture combinations, $F(3, 222) = 265.60, MSE = 353.59, p < .001, \eta_p^2 = .78$. As in Experiment 1, the first planned contrast revealed that individuals chose a specific picture more often after having seen the specific picture than after having seen a general picture in the presentation phase ($M = 81.53\%, SE = 1.73$ vs. $M = 18.30\%, SE = 1.55$; $F(1, 74) = 627.96, MSE = 7640.05, p < .001, \eta_p^2 = .90$). In sum, as in Experiment 1, the pattern of results was unsuited to disentangle the two-mental-representations view from the single-mental-representations view.

With regard to the sentence recognition forced-choice test, the repeated ANOVA revealed a significant effect for sentence-picture combinations, $F(3, 222) = 265.60, MSE = 353.59, p < .001, \eta_p^2 = .78$. As in Experiment 1, the first planned contrast revealed that individuals chose a specific sentence when they actually had seen general sentences/specific pictures than after having seen general sentences/general pictures, $F(1, 74) = 4.73, MSE = 537.81, p = .03, \eta_p^2 = .06$, supporting hypothesis 2.1 (i.e., the single-mental-representation view). The second planned contrast revealed no difference between specific sentences/general pictures and specific sentences/specific pictures, $F(1, 74) = 1.32, MSE = 577.06, p = .25, \eta_p^2 = .01$, contradicting hypothesis 2.2 (i.e., the two-mental-representations view) and therefore supporting again the single-mental-representation view, which predicted no difference. The third planned contrast confirmed that individuals chose a specific sentence more often after having seen the specific sentence than after having seen the general sentence in the presentation phase ($M = 75.58\%, SE = 2.08$ vs. $M = 14.47\%, SE = 1.63$; $F(1, 74) = 556.46, MSE = 8053.19, p < .001, \eta_p^2 = .88$).

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|l|l|}
\hline
 & GP/GS & GP/SS & SP/GS & SP/SS \\
\hline
Specific picture chosen (%) & 17.93 & 18.67 & 81.52 & 81.53 \\
(1.87) & (1.99) & (2.15) & (2.02) \\
Specific sentence chosen (%) & 11.56 & 73.98 & 17.38 & 77.17 \\
(1.35) & (2.47) & (2.42) & (2.52) \\
\hline
\end{tabular}
\caption{Means and standard errors (in parentheses) for choosing specific pictures and specific sentences as a function of condition in Experiment 2.}
\end{table}

Note. Gray shaded cells indicate false alarms. GP = general picture; GS = general sentence; SP = specific picture; SS = specific sentences.
3.3. Summary and discussion of Experiment 2

To summarize, the results of Experiment 2 showed the same pattern of results as Experiment 1: For sentence recognition, the single-mental-representation view was confirmed, indicating that specific picture information was integrated with general sentence information, leading to falsely choosing specific sentences; however, for picture recognition neither view was supported. Thus, although we excluded from analyses those items for which individuals showed a high performance level (i.e., ceiling effects), we could not distinguish between the two views regarding picture recognition. The findings will be discussed in the following section.

4. General discussion

The purpose of the present experiments was to shed light onto the kind of mental representations constructed under conditions of multimedia learning. More concretely, we tried to distinguish between the two-mental-representations view, which predicts that both a pictorial and a verbal mental representation are constructed and interconnected via referential connections, and the single-mental-representation view, which predicts that a single mental representation is constructed by merging information from the text and picture. To disentangle the two views, we adopted a paradigm used by Gentner and Loftus (1979), Pezdek (1977), and Wippich (1987): In the presentation phase, individuals were presented with four different text-picture combinations (general pictures/general sentences, general pictures/speciﬁc sentences, speciﬁc pictures/general sentences, and speciﬁc pictures/specific sentences). After the presentation phase, general and speciﬁc pictures (or sentences) were presented in a forced-choice format. Individuals were asked to decide which of the two items had been presented previously.

Across two experiments, individuals more often falsely decided in favor of the speciﬁc sentence when general sentences had been paired with speciﬁc pictures than with general pictures (hypothesis 2.1; see Fig. 1), indicating that they had integrated the speciﬁc picture information with the general sentence information. Moreover, no difference was observed when speciﬁc sentences had been paired with speciﬁc pictures or general pictures, contradicting the assumptions of the two-mental-representations view (hypothesis 2.2). Importantly, this latter pattern of results does not only contradict the two-mental-representations view, but corresponds to the single-mental-representation view, which presumed no differences.

On the other hand, regarding picture recognition, neither view was confirmed; that is, individuals did not more often falsely choose speciﬁc pictures when they had actually seen general pictures/speciﬁc sentences (contradicting hypothesis 1.1) and did not more often falsely choose speciﬁc pictures when they had actually seen speciﬁc pictures/general sentences (contradicting hypothesis 1.2).

One explanation for these data patterns might be that individuals in our study simply did not look at the pictures, because pictures were highly redundant to the sentences. However, this explanation is unlikely due to the fact that the combination of speciﬁc pictures and general sentences led to falsely choosing the speciﬁc sentences. This data pattern can only be explained by the fact that individuals had processed the pictures. Even more direct evidence that individuals processed the pictures comes from another study conducted in our lab in which we registered eye movement behavior of the participants during the presentation phase. Preliminary data analyses of that study conﬁrm our a priori assumption that individuals process both kinds of representations, irrespective of text-picture redundancy.

An alternative explanation might be that individuals relied on more perceptually-based representations when they had to decide which picture they had seen during presentation. In line with this reasoning, Radvansky and Copeland (2006) assumed that whereas mental models are created from a picture, they do not play a strong role during picture recognition, because the perceptually-based representations can be used more easily for picture recognition. Importantly, for sentences, no perceptually-based representations should be available. Therefore, individuals rely on their mental models in sentence recognition tasks, but rely on perceptually-based representations in picture recognition tasks.

Yet another explanation might lie in the different nature of text and pictures: Whereas pictures refer to a single situation and are thus highly speciﬁc, texts describe content in a less deterministic way (Stenning & Oberlander, 1995). For example, when reading about a “desk”, many different kinds of desks can be associated with this noun. On the other hand, when a picture depicts a desk, the kind of desk is determined because a concrete desk is depicted. Thus, imagine the case in which individuals see a general picture of a table paired with a speciﬁc sentence about a desk. When individuals have to decide which picture they had seen – the general picture with the table or the speciﬁc picture with the desk – they are less likely to recognize the picture with the desk because the depicted desk may not match the kind of desk they included in their mental model based on the noun “desk”, which was contained in the speciﬁc sentence. This in turn might explain why for picture recognition no effects were observed.

4.1. Implications

From a theoretical point of view, the observed pattern of results for sentence recognition speaks in favor of the single-mental-representation view. Thus, when individuals are presented with text and pictures, they construct a single mental representation containing the information from the text as well as the picture. Accordingly, the data for sentence recognition ﬁt well with the diagrammatic representation of cognitive theory of multimedia learning (e.g., Mayer, 2009, p. 79) and with the integrated model of text and picture comprehension (Schnozt & Bannert, 2003), which both posit that a single mental model is constructed containing information from text and pictures. On the other hand, the data ﬁt less well with the assumption of cognitive theory of multimedia learning that processing text and pictures results in two mental models interconnected via referential connections (Mayer, 1997). Thus, in our view the data are a ﬁrst step to clarifying the theoretical assumptions made by multimedia learning theories.

So far, the theoretical implications have been derived from the data pattern concerning sentence recognition. For picture recognition, on the other hand, no unambiguous data pattern was observed. Possible reasons were discussed above. It should be noted that these reasons may also have implications for both multimedia theories because neither of them would predict the observed data patterns for picture recognition. Instead, according to the cognitive theory of multimedia learning, the same results for sentence and picture recognition would have been expected. From the viewpoint of the integrated model of text and picture comprehension, the opposite pattern of results might be derived, namely, more pronounced differences for picture than for sentence recognition. This assumption can be derived from the fact that according to the theory, the mental model that combines information from text and picture has an analogous structure. Thus, the effects should be especially observable for picture recognition because pictures resemble the mental model more than sentences do. In sum, the fact that no effects were observed
for picture recognition contradicts both theories. In this respect, more research is needed to clarify the underlying reasons, which, in turn, might help in clarifying the assumptions made by multimedia theories.

4.2. Limitations and future research

Our studies differed from typical multimedia situations in several aspects: For example, in contrast to typical multimedia material, our materials were rather simplified in that we used short sentences and black-and-white drawings. Moreover, the presented text-picture stimuli were independent of each other; that is, a text-picture combination could be understood independently of the stimuli presented earlier. Furthermore, because we did not refer to existing systems or procedures in the real world but presented many different, highly familiar everyday situations, prior knowledge was assumed to be less important than in typical multimedia sessions. Thus, the experiments reported above led us to the conclusion that under multimedia conditions text and pictures are integrated into a single mental representation, at least when very basic, well controlled text-picture materials are used. A next step would be to investigate whether the observed findings can be replicated also with more complex text and pictures. This step is crucial to show that the mechanisms observed in the reported studies also apply to more naturalistic multimedia material.

Moreover, the materials used in the present experiments contained a large degree of information overlap between text and pictures; that is, they were redundant. Whether text-picture redundancy aids learning (e.g., Levie & Lentz, 1982; Mayer, 2009, p. 226; Paivio, 1990) or hampers learning (e.g., Chandler & Sweller, 1991; Sweller, van Merriënboer, & Paas, 1998) is still an open question. In the current studies, however, redundancy of text and pictures was a necessary prerequisite to allow for dual coding. Because redundancy was present in all four conditions, the effects of redundancy were not assumed to distort the results. Future research might investigate whether the degree of redundancy influences text-picture integration. Based on the observed data pattern, one might assume that text-picture representations might be most effective if text and pictures have a moderate to small degree of information overlap. Because the information from both external representations is merged within one mental representation, it might be advantageous to distribute the to-be-learned information across text and picture. On the other hand, presenting text and pictures with a large degree of overlap might not be more advantageous for learning than presenting only text or picture, when an adequate mental representation can be constructed based on one source of information alone. This would hold only for information that can be extracted from either text or picture with similar ease, that is, in cases where both representations are not only informationally, but also computationally equivalent (Larkin & Simon, 1987). When more complex information needs to be conveyed, it may nevertheless be apt to create redundant multimedia instruction, because different external representations may serve complementary functions (Ainsworth, 2006).

A potential methodological limitation lies in the forced-choice test applied after the presentation phase: First, a fixed set of pictorial and verbal items were used, without varying them between individuals. Second, the forced-choice recognition tests did not offer a third “I don’t know” option. Thus, individuals who forgot the picture or sentence were forced to decide in favor of one option. Therefore, future research should investigate whether the effects can be replicated with an improved version of the recognition test. Despite these limitations that can be well addressed in future research, the paradigm used in these experiments seems promising enough to open up several avenues for future research.

As discussed before, we were unable to observe the expected interaction for picture recognition. One explanation might be that with pictures, there is a perceptually-based representation that can be used by the individuals to solve the picture recognition task. Therefore, future studies should investigate the role of delayed testing on picture recognition, because it can be assumed that after some delay individuals rely more strongly on their mental models than on their perceptually-based (picture) representations (e.g., Shephard, 1967). Thus, after some delay effects for picture recognition could be expected. Importantly, in studies conducted in the context of multimedia research, the dependent variables are most often assessed immediately after learning. Therefore, if immediate testing is associated with more perceptually-based (picture) representations, these representations may also influence learning outcomes in typical multimedia studies. There are only a few multimedia studies that test performance after a delay: Interestingly, in these studies, it was reported that the results differed from the patterns observed immediately after presentation (e.g., Segers, Verhoeven, & Hulstijn-Hendrikse, 2008). Whether this was due to different mental representations being available with immediate versus delayed testing has to be clarified in further studies.

In addition, the paradigm is not only promising for examining the mental representations constructed during processing text and pictures, but it also allows for examining other research questions: for example, the effects of text and picture specificity on recognition performance. Although this was not the focus of the reported studies, the data indicate that individuals tended to falsely recognize specific pictures and general sentences. For example, the data in Table 1 regarding picture recognition (second row) indicate that individuals who had seen general pictures falsely recognized the specific picture in 22.07% of the cases. When looking at the false alarms for individuals who had seen specific pictures (by inverting the scores in Table 1), it becomes obvious that these individuals falsely recognized the general picture in only 12.82% of the cases. Thus, there seems to be a main effect of picture specificity, indicating that individuals tend to prefer specific pictures over more general pictures. When applying the same rationale to sentence recognition performance, it appears that individuals tend to prefer general over specific sentences. Thus, in sum, the paradigm also allows for analyzing general tendencies of individuals to select text or pictures varying in specificity.

To conclude, we think that the paradigm used in the reported studies is useful for investigating the mental representations resulting from processing text and pictures. As the construction of mental representations — measured via recall performance in the reported experiments — plays an important role in many learning situations, the results add to our understanding of the many different processes underlying learning with multimedia. Thus, these findings are highly relevant for refining existing cognitive theories of multimedia learning, from which practical recommendations regarding the design of multimedia instruction are derived. As has been emphasized by Mayer (2009) good theories of multimedia learning need to be grounded in cognitive research to allow for valid conclusions regarding not just which instructional variants yield better learning than others but also why they do so. In our view, knowing more about the memory representations underlying the integration process in multimedia learning and about ways of investigating them contributes to a better cognitive grounding of instructional research and thus should allow answering more “why” questions in the future.

Appendix

Examples of Stimulus Material.
<table>
<thead>
<tr>
<th>General Picture / Specific Sentence</th>
<th>General Picture / Specific Sentence</th>
<th>Specific Picture / General Sentence</th>
<th>Specific Picture / Specific Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Next to the house are parking places)</td>
<td>(Next to the row house are parking places)</td>
<td>(Next to the house are parking places)</td>
<td>(Next to the row house are parking places)</td>
</tr>
<tr>
<td>(On the small island is only a tower)</td>
<td>(On the small island is only a lighthouse)</td>
<td>(On the small island is only a tower)</td>
<td>(On the small island is only a lighthouse)</td>
</tr>
<tr>
<td>(The big fish swims among the small fishes)</td>
<td>(The big shark swims among the small fishes)</td>
<td>(The big fish swims among the small fishes)</td>
<td>(The big shark swims among the small fishes)</td>
</tr>
<tr>
<td>(In the room lies a fringe carpet)</td>
<td>(In the living room lies a fringe carpet)</td>
<td>(In the room lies a fringe carpet)</td>
<td>(In the living room lies a fringe carpet)</td>
</tr>
</tbody>
</table>

**Note.** Translations in English were not given in the original experiment.

**References**


