Running head: INTEGRATED MODEL OF TEXT AND PICTURE COMPREHENSION

Chapter 4

Integrated Model of Text and Picture Comprehension

Wolfgang Schnotz

University of Koblenz-Landau, Germany

To appear in:

R.E. Mayer (Ed.), The Cambridge Handbook of Multimedia Learning. Cambridge:

CambridgeUniversity Press. (2nd Edition)

Date submitted: March 10, 2013

Date accepted: March 22, 2013

Abstract

This chapter presents an integrated model of text and picture comprehension that takes into account that learners can use multiple sensory modalities combined with different forms of representation. The model encompasses listening comprehension, reading comprehension, visual picture comprehension and auditory picture comprehension (i.e., sound comprehension). The model's cognitive architecture consists of modality-specific sensory registers, working memory and long term memory. Within this architecture, a distinction is made between perception-bound processing of text surface or picture surface structures, on the one hand, and cognitive processing of semantic deep structures, on the other hand. The perception-bound processing of text surface structures includes phonological and graphemic input analyses with graphemic-phonemic conversion, leading to lexical patterns. The perception-based processing of picture surface structures includes visual or acoustic non-verbal feature analyses leading to visuo-spatial patterns or sound patterns. The cognitive processing includes descriptive processing of lexical patterns (via parsing) which leads to propositional representations, and depictive processing of spatial or sound patterns (via structure mapping) which leads to mental models. Propositional representations and mental models interact via model construction and model inspection processes. After presenting the integrated model of text and picture comprehension, the chapter derives predictions, which can be empirically tested. It reports research findings that can be explained by the model, and it derives practical suggestions for instructional design. Finally, the chapter discusses limitations of the model and points out directions for further research.

The term *multimedia* has different meanings at different levels. At the level of technology, it means the use of multiple delivery media such as computers, screens, and loudspeakers. At the level of presentation formats, it means the use of different forms of representation such as texts and pictures. At the level of sensory modalities, it means the use of multiple sensory organs such as the eye and the ear. Although highly important in terms of practical reliability, the level of technology is not very interesting from a psychological point of view: Comprehending a text printed on paper does not fundamentally differ from comprehending the same text on a computer screen. In fact, comprehension is highly dependent on what kind of information is presented and how it is presented. The psychology of multimedia learning focuses therefore on the level of presentation formats and on the level of sensory modalities.

What is multimedia learning? From a psychological point of view, the core of multimedia learning is the combined comprehension of text and pictures (Mayer, 1997). This does not necessarily require high-technology. Multimedia learning is also possible with printed books or blackboards instead of computer screens and with the human voice instead of loudspeakers. Multimedia learning is therefore not a modern phenomenon. Instead, it has a long tradition going back to Comenius (1999) who emphasized the importance of adding pictures to texts in his pioneer work *Orbis sensualium pictus* (published first in 1658).

Multimedia learning can occur in different forms. A learner can listen to a lecture accompanied by pictures (i.e., lecture-based multimedia learning). He/she can read a book with pictures (i.e., book-based multimedia learning). Finally, he/she can read an illustrated text from the internet on a computer screen or listen to a text accompanied by pictures from a loudspeaker (i.e., computer-based multimedia learning) (Mayer, 2009).

Individuals usually combine these different kinds of multimedia learning. Consider the following example. A teacher explains to her class of 8th-graders the migration of birds in Europe. She presents a map of the continent (shown in Figure 4.1(a)), which indicates where some birds live in summer and where they stay in winter. While pointing on the map, she gives oral explanations like the following:

(a) *"Many birds breed in Middle and Northern Europe in summer, but do not stay there during winter. Instead, they fly in September to warmer areas in the Mediterranean area. These birds are called migrant."*

Figures 4.1(a-c) about here

At the end of the lesson, Daniel, one of her students, has to learn as a homework task about a specific bird, the marsh harrier, and to give a report to his classmates the next day. Daniel walks into a library and looks up in a printed encyclopedia of biology, where he finds a drawing of the marsh harrier (shown in Figure 4.1(b)) and the following text:

(b) *"The marsh harrier is bird of prey with an average wingspan of 47" and a face similar to owls. The drawing shows the typical gliding position of the bird. The marsh harrier is usually found in wetlands, especially in marshes, swamps, and lagoons. It feeds mostly on small birds or mammals (like rodents or rabbits), and on reptiles. The marsh harrier is migrant."*

As the encyclopedia does not contain further information about the bird's migration, Daniel decides to search in the Internet, where he finds a web site including a bar graph (shown in Figure 4.1(c)) and the following text:

(c) *"The marsh harrier is found all year round in Spain, France and around the Mediterranean. In other areas of Europe the bird is migrant, breeding in Middle and Northern Europe while wintering in tropical marshes and swamps in North Africa. The bar graph shows a typical frequency pattern of marsh harriers in a Middle European habitat."*

Furthermore, the website offers a sound button. After clicking on it, Daniel hears the typical call of a marsh harrier near its breeding place.

Altogether, Daniel has practiced three kinds of multimedia learning using various external sources of information. At school, he has performed lecture-based multimedia learning, using the map and the teacher's oral text as information sources. In the library, he has performed book-based multimedia learning, using the drawing of the bird and the printed text as information sources. With the Internet, he has performed computer-based multimedia learning, using the bar graph, the on-screen text and the sound pattern as information sources. In each case, information was presented to him in different formats such as visual texts, visual pictures (a map, a drawing, a bar graph) and sound, and he has processed information through different sensory modalities: the visual modality (written text and pictures) and the auditory modality (oral text and sound).

As the example demonstrates, multimedia learning environments can be rather complex and they can involve a variety of external representations of the learning content. These representations can take different forms such as spoken text, written text, maps, drawings, graphs, and sound. Multimedia learning occurs when an individual understands what is presented, that is, when he/she uses the external representations in order to construct internal (mental) representations of the learning content in working memory and if he/she stores these representations in long-term memory.

In the first part of this chapter, a distinction between two different forms of representations is made and applied to both external and internal representations. The second part investigates how multimedia comprehension and learning is constrained by the human cognitive architecture. In the third part, the theoretical concepts introduced before will be combined into an integrated model of text and picture comprehension, which involves listening comprehension, reading comprehension, visual picture comprehension, and auditory picture comprehension (i.e. sound comprehension). The fourth part presents empirical evidence for the integrated model, whereas the fifth part explains what kind of instructional consequences can be derived from the integrated model. Finally, the sixth part points out limitations of the model and suggests directions of future research in the area.

External and Internal Representations

Forms of Representation

How many basic forms of representation exist? Despite of numerous variants of representations, there are only two basic forms of representations: descriptions and depictions. Texts are the most common kind of descriptions. However, there are also other kinds of descriptive representations. Mathematical expressions such as $V=s^3$ (describing the relation between a cube's size and its volume) or the formula $F=m*a$ in physics (describing the relation between force, mass and acceleration according to Newton's Second Law) are also descriptive representations. Descriptive representations consist of symbols. Symbols are signs that have no similarity with their referent (Peirce, 1931/1958). The word 'bird', for example, has no similarity with a real bird. It is a symbol, and its meaning is based on a convention. In texts, we use nouns (such as 'bird' and 'breeding') as symbols for objects and events. We use verbs and prepositions (such as 'feed' and 'on') as symbols for relations, and we use adjectives (such as 'small' and 'migrant') as symbols for attributes.

Pictures such as photographs, drawings, paintings and maps are depictive representations. It should be noted, however, that pictures are not the only kind of depictive representations. A miniature model of a building, a line graph, or the swing of a measuring tool pointer are also depictive representations. Depictive representations consist of icons. Icons are signs that are associated with their referent by similarity or by another structural commonality. A map such as those in Figure 4.1(a) or the drawing of a bird as those in Figure 4.1(b) are graphical objects that have some similarity with the corresponding referent (i.e., the European continent or the marsh harrier). Graphs have a more abstract structural commonality with their referent. The meaning of the bar graph shown in Figure 4.1(c), for example, is based on an analogy: The height of the bars corresponds to the number of marsh harriers observed in a habitat during the corresponding month, and the sequence of bars corresponds to the sequence of months during the year.

Descriptive representations and depictive representations have different uses for different purposes. On the one hand, descriptive representations are more powerful in expressing abstract knowledge. For example, it is no problem to say '*The Marsh Harrier feeds on mammals or reptiles',* which connects abstract concepts (e.g. *'mammals'*, *'reptiles'*) by a disjunctive *'or'*. In a depictive representation, on the contrary, it is only possible to show a specific mammal (e.g. a mouse) or a specific reptile (e.g. a lizard). The disjunctive 'or' cannot be represented by only one picture. It requires a series of pictures (e.g. one picture showing the bird eating a mouse and another picture showing the bird eating a lizard). On the other hand, depictive representations have the advantage of being informationally complete. A map, for example, includes all geometric information of the depicted geographical area, and a picture of a marsh harrier eating a mouse includes not only information about the shape of the bird and the shape of a mouse, but necessarily also information about their size, about their orientation in space, how it holds its pray, etc. Depictive representations are therefore more

useful to draw inferences, because the new information can be read off directly from the representation (Kosslyn, 1994).

Mental Representations

Does the distinction between descriptive and depictive representations apply also to internal (i.e. mental) representations? Research on text comprehension suggests that learners reading a text or listening to a text construct three kinds of mental representations (Graesser, Millis & Zwaan, 1997; Kintsch, 1998; McNamara, 2007; van Dijk & Kintsch, 1983; van Oostendorp & Goldman, 1999; Weaver III, Mannes & Fletcher, 1995). For example, when a learner reads a sentence like, *Some migrant birds fly to the South of Europe for wintering*, he/she forms a mental representation of the text-surface structure. This text-surface representation cannot be referred to as understanding yet, but it allows repetition of what has been read. Based on this surface representation, the reader then constructs a propositional representation. This representation includes the ideas expressed in the text at a conceptual level, which is independent from the specific wording and syntax of the sentence. In the previous example, this would include the idea that migrant birds in Europe fly to the South in September, represented by the proposition, FLY(agent: MIGRANT BIRDS, location: EUROPE, aim: SOUTH, time: SEPTEMBER). Finally the reader constructs a mental model of the text content. In the previous example, this could be a mental map of Europe including a movement from the North to the South.

Research on picture comprehension suggests that when learners understand a picture, they also construct multiple mental representations (Kosslyn, 1994; Lowe, 1996). Accordingly, a learner creates a perceptual representation (i.e. a visual image) of the picture, and he/she constructs then a mental model of the picture's content. For example, when a learner understands the bar graph shown in Figure 4.1(c), he/she perceives vertical bars on a

horizontal line and creates a corresponding visual image. Based on this visual image, he/she constructs a mental model of a Middle European habitat that includes different numbers of marsh harriers during the course of the year. The mental model can be used for reading off specific information as, for example, that the birds stay in this habitat during the summer. The information read-off from the model is again encoded in a propositional format such as, for example, STAY(agent: BIRDS; location: HABITAT, time: SUMMER).

The distinction between descriptive and depictive representations mentioned above applies also to these mental representations. A text surface representation and a propositional representation are descriptive representations, as they use symbols in order to describe a subject matter. A visual image and a mental model, on the contrary, are depictive representations, as they are assumed to have an inherent structure that corresponds to the structure of the subject matter (Johnson-Laird, 1983; Kosslyn, 1994). A visual image is sensory specific, because it is linked to the visual modality, whereas a mental model is not sensory specific because it is able to integrate information from different sensory modalities. It is possible, for example, to construct a mental model of some spatial configuration based on visual, auditory and touch information. This implies that a mental model is a more abstract than a visual image. In picture comprehension, mental models and visual images can also differ in terms of their information content. On the one hand, irrelevant details of the picture, which are included in the visual image, may be ignored in the mental model. On the other hand, the mental model contains additional information from prior knowledge that is not included in the visual image. In understanding bird migration, for example, a mental model of the European continent might include snowfall in Northern areas during winter, although no snow is indicated in the map.

Based on the distinction between descriptive and depictive representations, Schnotz and Bannert (2003) have proposed a theoretical framework for analyzing text and picture

comprehension. The framework, which is shown in Figure 4.2, includes a branch of descriptive representations (left side) and a branch of depictive representations (right side) with correspondingly different types of information processing. The descriptive branch involves the external text, the mental text surface representation and the mental propositional representation of the subject matter. Information processing in the descriptive branch implies (sub-semantic and semantic) analysis of symbol structures. The depictive branch involves the external picture, the visual image of the picture and the mental model of the subject matter. Information processing in the depictive branch implies analog structure-mapping (based on perception and thematic selection). The framework corresponds to the dual-coding concept of Paivio (1986), who assumes a verbal system and an image system in the human mind with different forms of mental codes. However, contrary to the traditional dual-coding theory, the framework assumes that multiple representations are formed in text comprehension as well as in picture comprehension.

Figure 4.2 about here

Cognitive Architecture for Text and Picture Comprehension

When learners understand texts and pictures, they construct multiple mental representations in their cognitive system. Research in cognitive psychology suggests that the architecture of the human cognitive includes multiple memory systems. A common view proposed by Atkinson and Shiffrin (1971) distinguishes three memory sub-systems -- sensory registers, working memory, and long-term memory -- with different functions and different constraints on processing texts and pictures.

Sensory Registers

Information enters the cognitive system from the outside world through sensory organs, which convey the information through sensory channels to working memory. It should be noted that there is no inherent relationship between sensory modalities and representational formats. For example, written text is usually visual language read with the eyes, but can also be read with the fingers (e.g. in the case of blind people reading Braille). Similarly, pictures are usually seen with the eyes, but can sometimes also be perceived by touch (e.g. maps for blind people). Spoken text is usually perceived by the ear, but deaf people can also read lips and touch the vibration of the larynx. Auditory pictures (i.e. sound patterns imitating an original sound as, for example, the call of a bird) are perceived by the ear too. Although there are multiple sensory modalities which can be involved in text and picture comprehension, we will consider in the following only the visual and the auditory modality.

Visual information that meets the eye is stored very briefly (i.e. for less than 1 second) in a visual register. If attention is directed to information in the visual register, the information gets transmitted to visual working memory. Auditory information that meets the ear is stored briefly (i.e. for less than 3 seconds) in an auditory register. If attention is directed to information in the auditory register, the information gets transmitted to auditory working memory.

Working Memory

Written or spoken text and visual or auditory pictures are further processed in a working memory with a highly constrained capacity for storing and processing of information (see Chapter 25). According to Baddeley (1986), working memory consists of a central executive and different subsystems for the storage of information.

Two of these subsystems have received much attention in research: auditory working memory and visual working memory. Auditory working memory is conceived as a phonological-articulatory loop. Visual working memory is conceived as a visuo-spatial sketchpad. The phonological-articulatory loop specializes on verbal material presented in auditory modality, but can also deal with non-verbal sound. It has limited capacity corresponding on average to what can be articulated within about 2 seconds. Nevertheless, people with a highly reduced phonological-articulatory loop are still capable of normal language comprehension (Vallar & Shallice, 1990; Baddeley, 2000). Spoken text activates phonological lexical patterns, whereas auditory pictures activate sound patterns in auditory working memory. The visuo-spatial sketchpad specializes on spatial information presented in visual modality. It has a limited capacity of about 5 units on the average. Written text activates graphemic lexical patterns, whereas visual pictures activate visuo-spatial patterns in visual working memory.

As working memory plays an important role at higher levels of text comprehension too (Daneman & Carpenter, 1983), one can furthermore assume a propositional subsystem that allows holding a limited number of propositions simultaneously in working memory (Kintsch & van Dijk, 1978). Propositions result from descriptive processing of phonological or graphemic lexical patterns through parsing the incoming word sequences combined with prior knowledge. Finally, research findings suggest a subsystem for mental model construction in working memory. Mental model construction seems to be influenced by the visuo-spatial sketchpad rather than the phonological-articulatory loop (Friedman & Miyake, 2000). More specifically, it is highly related to spatial cognitive processing (Sims & Hegarty, 1997). Corresponding to these findings, research by Knauff and Johnson-Laird (2002) indicates that visual imagery and spatial reasoning are based on different cognitive subsystems. This suggests a distinction between a visual working memory (or sketchpad) for visual images and

a spatial working memory for mental model construction. Accordingly, mental models result from depictive processing of visuo-spatial or sound patterns through structure mapping.

Long Term Memory

Text comprehension and picture comprehension requires prior knowledge stored in long-term memory, which includes lexical knowledge as well as perceptual and cognitive world knowledge. Lexical knowledge encompasses the mental phonological lexicon and the mental graphemic lexicon, which include knowledge about auditory or visual word forms. The phonological lexicon (also called 'auditory lexicon') includes phonological lexical patterns, which represent knowledge about the sound of spoken words required for spoken word recognition. Listening to a text implies activation of such phonological lexical patterns in working memory. Individuals who suffer from word deafness (due to brain injuries) have a deficient phonological lexicon: They can hear sounds, but cannot separate and identify words when listening to spoken language. Individuals who suffer from word meaning deafness can repeat spoken words without understanding them, although they can understand written words. These individuals possess a phonological lexicon, but this is unconnected to semantic (long-term) memory. The graphemic lexicon (also called 'visual' or 'orthographic lexicon') includes graphemic lexical patterns, which represent knowledge about the visual appearance of written words required for written word recognition. Reading a text implies activation of such graphemic lexical patterns in working memory. Individuals who suffer from pure alexia (due to illiteracy or brain injuries) have a deficient graphemic lexicon: They can understand spoken words, but cannot understand written words although their vision is intact (Ellis & Young, 1996).

Perceptual world knowledge refers to the appearance of objects as, for example, how different kinds of birds typically look like. This knowledge is needed for the visual perception

or imagination of objects, that is, for the creation of corresponding visuo-spatial patterns in working memory (Kosslyn, 1994; Rosch, 1978). Objects can be recognized faster and more easily, when they are presented from a typical perspective (such as the bird shown in Figure 4.1(b)) than when they are presented from an unusual perspective (Palmer, Rosch & Chase, 1981). Conceptual world knowledge refers to the relations within a domain as, for example, the breeding of birds and the meteorological conditions in different seasons. This knowledge is needed both for the construction of a propositional representation and the construction of a mental model (e.g. of bird migration) in working memory.

Text and picture comprehension are therefore not only based on external sources of information (i.e. the text and the picture), but also on prior knowledge as an internal source of information. Prior knowledge can partially compensate for a lack of external information, for lower working memory capacity (Adams, Bell & Perfetti, 1995; Miller & Stine-Morrow, 1998), and for deficits of the propositional representation (Dutke, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996; Soederberg Miller, 2001). There seems to be a trade-off between the use of external and internal information sources: Pictures are analyzed more intensively if the content is difficult and the learners' prior knowledge is low (Carney & Levin, 2002).

Integrated Comprehension of Text and Pictures

The idea of a cognitive architecture including multiple memory systems with multiple sensory channels of limited capacity and a working memory of limited capacity operating on descriptive and depictive representations are combined in the following into an integrative model of text and picture comprehension (or ITPC model for short). The model integrates the concepts of multiple memory systems (Atkinson & Shiffrin, 1971), working memory (Baddeley, 1986, 2000), and dual coding (Paivio, 1986). It furthermore integrates the idea of multiple forms of mental representations in text comprehension or picture comprehension

(Kosslyn, 1994; van Dijk & Kintsch, 1983) and neuropsychological models of word recognition and reading (Ellis & Young, 1996). Naturally, the model has commonalities with the precursive model of text and picture comprehension of Schnotz and Bannert (2003) and it has commonalities with the cognitive theory of multimedia learning (CTML) of Mayer (2009; see also Chapter 3). The model, which is schematically shown in Figure 4.3, aims at representing the single or combined comprehension of spoken text, written text, visual pictures and auditory pictures (i.e. sound images). It is based on the following assumptions:

- Text and picture comprehension take place in a cognitive architecture including modality-specific sensory registers as information input systems, a working memory of limited capacity, and a long-term memory.
- Verbal information (i.e. information from spoken or written texts) and pictorial information (i.e. information from visual pictures or from sound pictures) is transmitted to working memory through visual channels and auditory channels. The channels have limited capacity to process and transmit information.
- Further semantic processing in working memory takes place in two different subsystems: a descriptive subsystem and a depictive subsystem. Text (spoken or written) is first processed in the descriptive subsystem followed by the depictive subsystem. Pictures (visual or auditory) are first processed in the depictive subsystem followed by the descriptive subsystem.
- Text and picture comprehension are active processes of coherence formation. In comprehension, individuals engage in building coherent knowledge structures from the available external verbal and pictorial information and from their prior knowledge.

Figure 4.3 about here

A distinction between perceptual surface structure processing and semantic deep structure processing can be made within the model. Perceptual surface structure processing refers to the information transfer from the surface structure of texts and pictures to working memory. It is characterized by (verbal) phonological or graphemic analyses and (non-verbal) visual or acoustic feature analyses leading to corresponding input patterns in auditory or visual working memory. Semantic deep structure processing refers to the cognitive processing within working memory which results in propositional representations and mental models as well as the information exchange between long-term and working memory. It is characterized by the functioning of the descriptive and the depictive subsystem and their interaction.

Listening comprehension. If a spoken text is understood, auditory verbal information enters the auditory register through the ear and is then object of phonological input analysis which identifies phonemes within the acoustic input leading to phonological lexical patterns. Further descriptive processing (parsing of word sequences and further semantic analysis) leads to a propositional representation, which finally triggers the construction or elaboration of a mental model. In the example of a text on bird migration, phonological analysis of the spoken word *bird* leads (via the mental lexicon) to the activation of its phonological pattern in auditory working memory. Further processing through the descriptive subsystem results in the activation of the concept BIRD, which is then included into a propositional representation. This representation finally triggers the construction of a mental model of bird migration.

Reading comprehension. If a written text is understood, visually presented verbal information enters the visual register through the eye and is then is subjected to graphemic

input analysis, which identifies graphemes within the visual input. In skilled reading, this analysis leads to graphemic lexical patterns. These patterns are further processed in the descriptive subsystem. This results in the formation of a propositional representation, which in turn triggers the construction or elaboration of a mental model. In a text on bird migration, for example, graphemic analysis of the written word *bird* leads (via the mental lexicon) to the activation of its graphemic pattern in visual working memory. Further processing through the descriptive subsystem results in the activation of the concept BIRD, which is included into a propositional representation. This representation finally triggers the construction of a mental model of bird migration.

In non-skilled reading due to a deficient graphemic mental lexicon (e.g. by reading beginners), the individual has to apply grapheme-phoneme conversion rules by engaging in tedious phonological recoding of the visual input which finally allows understanding of the internally spoken text. The grapheme-phoneme conversion rules, which are neither lexical nor semantic, convert letter strings into phoneme strings (Coltheart, Rastle, Perry, Langdon & Ziegler, 2001). With a comprehensive graphemic lexicon, on the contrary, texts can be understood via the activation of graphemic lexical patterns without inclusion of any acoustic patterns (Ellis & Young, 1996, p. 219). Nevertheless, even skilled readers engage at least to some extent in graphemic-phonemic lexical conversion (Rieben & Perfetti, 1991) operating at the whole-word (lexical) level instead of the (sub-lexical) grapheme-phoneme level. It should be noted that this conversion represents per se a non-semantic lexical route of word recognition: it allows word recognition even without understanding the meaning of the word (Coltheart, Rastle, Perry, Langdon & Ziegler, 2001).

Thus, when familiar written words are recognized, the activated graphemic lexical patterns usually activate also phonological lexical output patterns which allow the reader to pronounce these words. The pronounciation does not imply reading aloud. It can also take the

form of *inner speech* as a hidden phonological output that can be heard by the reader through his/her *inner ear*. The inner pronunciation feeds into the phonological input analysis and activates phonological lexical input patterns, which are further processed through the descriptive subsystem as mentioned above. As a result, read words can be processed both via graphemic and phonological lexical patterns (Ellis & Young, 1996, p. 219). Graphemic-phonemic lexical conversion seems to be especially important to trigger parsing procedures (e.g. analyzing word order in sentence structure analysis). Caramazza, Berndt and Basili (1983) found that anomalies in the syntax of sentences are more easily detected when inner speech is possible than when it is suppressed. Although there is a direct route in reading from graphemic lexical patterns to further semantic analysis in the descriptive subsystem, this route does obviously not trigger syntactic analysis of the sentences. Syntactic processes "… appear to operate upon a speech-based code, so that written sentences which are to undergo syntactic analysis must first be converted into spoken form and then recycled back to auditory comprehension processes" (Ellis & Young, 1996, p. 221).

Visual picture comprehension. If a visual picture is understood, visual pictorial information enters the visual register through the eye and is then subjected to visual feature analysis which results in visuo-spatial patterns in working memory as a visual perceptual representation of the picture. Further depictive processing through the mapping of selected perceptual structures leads to the construction or elaboration of a corresponding mental model. This model can then be used by model inspection for reading off new information which is encoded in a propositional format in working memory. For example, if a map about bird migration in Europe such as in Figure 4.1(b) is understood, the visual pattern of the map creates via visual feature analysis an internal visual image of the map in visual working memory. Selected information is then further processed via structure mapping, which results in the construction or elaboration of a mental model of bird migration in Europe. The

individual can then read off further information from the model (such as the fact that migrant birds fly from Northern Europe to the Mediterranian area in fall).

Auditory picture comprehension (sound comprehension). If a sound is understood, auditory pictorial information enters the auditory register through the ear and is then object of acoustic feature analysis which results in sound patterns in working memory as an auditory perceptual representation of the sound. Further depictive processing through the mapping of selected perceptual structures leads to the construction or elaboration of a corresponding mental model. This model can then be used by model inspection for reading off new information which is encoded in a propositional format in working memory. For example, if the call of a marsh harrier (as bird of prey) and a call of a small bird (as its possible prey) are heard, acoustic feature analysis leads to sound patterns forming an auditory perception (i.e. an auditory internal image) in auditory working memory. If an individual has sufficient knowledge about different birds, selected information can be further processed via structure mapping which leads to the construction or elaboration of the mental model of a predator-prey scenario. The individual can then read off further information from the mental model (e.g. that a small bird is in danger of falling a prey to a marsh harrier).

It should be noted that according to this theoretical model, picture comprehension provides a more direct access to mental model construction than text comprehension, because pictures are immediately processed by the depictive subsystem, whereas texts are first processed by the descriptive subsystem which usually leaves some ambiguity to be removed in the following by the depictive subsystem (cf. Ainsworth, 1999)

Which cognitive processes lead to meaningful learning? Meaningful learning from text and pictures requires a coordinated set of cognitive processes including selection of information, organization of information, activation of prior knowledge, and active coherence

formation by integration of information from different sources. In comprehension of written or spoken texts, the learner selects relevant verbal information from words, sentences and paragraphs as an external source of information. He/she organizes the information, activates related prior knowledge as an internal source of information, and constructs both a coherent propositional representation and a coherent mental model. In comprehension of visual pictures, the learner selects relevant pictorial information from a drawing, a map or a graph as an external source of information, organizes the information, activates related prior knowledge as a further source of information and constructs a coherent mental model complemented by a propositional representation. In comprehension of auditory pictures (sound comprehension), the learner selects relevant acoustic information, organizes the information, activates related prior knowledge as internal source of information and constructs a coherent mental model complemented by a propositional representation.

As has been shown above, the ITPC model is embedded into a broader framework of human cognition which incorporates:

- concepts from semiotics (distinguishing between symbols and icons, or descriptions and depictions, respectively),
- concepts from text processing research (distinguishing between text surface representations, propositional representations, and mental models),
- concepts from picture processing (distinguishing between visual imagery and mental models),
- concepts from cognitive neuropsychology (distinguishing between phonological and graphemic mental lexicons as well as different kinds of graphemic-phonemic conversion),

- concepts from memory research combined with general ideas on the human cognitive architecture (multiple memory stores including the sub-structure of working memory).

Furthermore, the ITPC model takes the active and constructive nature of comprehension and learning into account. Most importantly, the model offers a framework for the analysis of text and picture comprehension that allows explanation of a broad variety of empirical findings.

Empirical Evidence

In order to demonstrate its validity, the ITPC model should be able to predict under which conditions combinations of text and pictures will be beneficial for learning. However, the model should also be able to predict under which conditions such combinations will have detrimental effects. This part of the chapter analyzes how far the ITPC model is able to successfully predict or explain positive and negative effects of using texts and pictures instead of using texts alone or pictures alone.

Positive Effects of Combining Texts and Pictures

Numerous studies have shown that students usually learn better from words and pictures than from words alone (Levie & Lentz, 1982; Levin, Anglin, & Carney, 1987). This is what Mayer (1997) has called the *multimedia effect* (see Chapter 7). The effect is bound to specific conditions.

Reading skills and prior knowledge. The ITPC model considers text comprehension and picture comprehension as different routes of constructing mental models and propositional representations using prior knowledge as a third source of information. If one route does not work well or if one source provides only little information, the other sources and routes become more important. When learners are poor readers, picture comprehension becomes more important. Thus, the ITPC model predicts that poor readers profit more from

illustrations in written texts than good readers. This prediction corresponds to various empirical findings reported by Cooney and Swanson (1987), Levie and Lentz (1982), and by Mastropieri and Scruggs (1989).

As text comprehension and picture comprehension are considered as different routes to the construction of mental representations, the ITPC model also implies the possibility that one route replaces the other one to some extent: Pictures can be used instead of a text, and a text can be used instead of pictures. The model therefore predicts that if a picture is added to a text and if the same amount of mental effort is invested into learning, text information becomes less important due to the additional picture information. The text will therefore be processed less deeply, resulting in lower memory for text information than if the text had been processed without pictures. Corresponding findings have been reported by Mayer and Gallini (1990) and by Schnotz and Bannert (1999).

When learners have low prior knowledge, they possess only a poor internal source of information. Mental model construction only from written text can become too difficult under these conditions. Adding a picture as another source of information can then considerably enhance comprehension, because it offers an additional route for mental model construction. Learners with high prior knowledge, on the contrary, are able to construct a mental model also without pictorial support. The integrated model therefore predicts that learners with low prior knowledge profit more from pictures in texts than learners with high prior knowledge. This corresponds to the results of various studies which found that pictures in texts are more beneficial for students with low prior knowledge than for those with high prior knowledge (Mayer, 2009; see also Chapter 24).

Redundancy. Contrary to the dual coding theory, which assumes that adding pictures to texts leads always to better learning, because two codes in memory are better than one, the

ITPC model predicts that the combination of texts and pictures can have also detrimental effects, because high prior knowledge can suspend the multimedia effect. Learners with high prior knowledge frequently do not need both text and pictures as information sources, because one source provides all information required for mental model construction. In this case, adding a picture to a written text means adding redundant, unneeded information. Although one of the two information sources is not needed, the eye wanders between the two sources, which implies split of attention. Thus, the learner loses time and mental effort with search for redundant information without a benefit for learning. This negative effect has been called the *redundancy effect* (Chandler & Sweller, 1996; Sweller, van Merrienboër & Paas, 1998; see also Chapter 10). This effect implies that a combination of text and pictures which has a positive effect on mental model construction when learners have low prior knowledge may have a negative effect on learning when prior knowledge is high. Experts possibly perform better with only one information source (i.e. text or picture) instead of two information sources (i.e. text and pictures). Corresponding findings have been reported by Kalyuga, Chandler and Sweller (2000), who have named this the *expertise reversal effect*.

Coherence and contiguity. Students learn better from words and pictures than from words alone, if the words and pictures are semantically related to each other (the *coherence* condition) and if they are presented closely together in space or in time (the *contiguity* condition). These findings are explained by the ITPC model in a similar way as by the CTML of Mayer (2009). The ITPC model assumes that a text and a picture can only contribute to joint mental model construction if the text and the picture are semantically related. This corresponds to the *coherence* condition. The model further assumes that text and picture can only contribute to joint mental model construction if corresponding text information and picture information are simultaneously available in working memory. As information decays

quickly from working memory, this requires combined presentation of words and pictures as far as possible. This corresponds to the *contiguity* condition (see Chapter 13).

If a picture is combined with written text, all information has to enter working memory through the visual register. The eye has to switch between pictures and words (i.e. between visual non-verbal feature analysis and graphemic input analysis) so that only one kind of information can be processed at the same time. This *split attention* implies unproductive search processes from the picture to the text and vice versa, and it affects the simultaneous availability of verbal and pictorial information in working memory (see Chapter 8). When pictures and related written words are presented closely to each other (i.e. *spatial contiguity*), visual search processes are reduced. Spatial contiguity is a way to minimize the loss of information due to split of attention and to allow an approximately simultaneous availability of pictorial and verbal information in working memory. In other words, spatial contiguity is means to maximize temporal contiguity in working memory under the condition of a picture with written text. Fully simultaneous availability, however, can only be ensured when a picture is combined with auditory text because pictorial and verbal information can then be processed at the same time (i.e. *temporal contiguity*) and be kept simultaneously in working memory. In this case, no split of attention is required because the learner can devote his/her full visual attention to the picture and his/her full auditory attention to the text (Mousavi, Low & Sweller, 1995). This has led to the assumption of a modality effect.

Modality. The modality effect states that students learn better from multimedia instructional messages when text is spoken rather than written (Mayer & Moreno, 1998; Moreno & Mayer, 1999; Ginns, 2005; see also Chapter 9). The modality effect is a derivate of the multimedia effect, because the rationale behind the modality effect is to take full advantage of text-picture combinations (i.e., of the multimedia effect) by maximizing contiguity of verbal and pictorial information or by minimizing any obstacles for

simultaneous availability of verbal and pictorial information in working memory, respectively. The key to minimizing the obstacles and to maximizing contiguity is the combination of auditory presentation of text and visual presentation of pictures. As the modality effect is a derivate of the multimedia effect, a modality effect is only to be expected if there is also a multimedia effect. If there is no multimedia effect, no modality effect is to be expected either.

Currently, there is no straightforward answer to the question where the modality effect comes from. The most popular explanation is the avoidance of split attention as mentioned above (Leahy, Chandler & Sweller, 2003; Mayer & Moreno, 1998; Mousavi, Low & Sweller, 1995). Split attention is indeed a fundamental problem when written text is combined with animation: As soon as the learner reads some text, he/she is at risk of missing important pictorial information, which can be avoided by using spoken text. Besides split attention, Moreno and Mayer (1999) have argued for an additional explanation of the modality effect. They presented text and pictures to learners in a consecutive way and, thus, avoided split attention. Nevertheless, spoken text with pictures resulted in better learning than written text with pictures. The authors argued that part of the modality effect results from the amount of involved working memory capacity. Text and picture comprehension are enhanced, if both visual memory and auditory working memory are involved, even if the two systems receive their input only in a consecutive manner. Although this explanation seems to be plausible, the ITPC model does not support this assumption because both comprehension of spoken and comprehension of written text involve auditory working memory. Research findings suggest that even experienced readers engage in graphemic-phonemic lexical conversion and recode at least parts of the visual information into auditory information (Ellis & Young, 1996; Rieben & Perfetti, 1991). Similarly, Baddeley (1999) assumes that verbal information – either spoken or written – is generally processed in the phonological loop rather than the visuo-spatial

sketchpad. Rummer, Schweppe, Fürstenberg, Seufert & Brünken, 2010) have suggested an *auditory-recency explanation* of the modality effect when text material consists of single sentences presented alternately with pictures. The authors argue that due to the longer duration of acoustic information in the auditory register compared to the visual register, a sentence can be better maintained in working memory after having heard than after having read it. Last but not least, a modality effect can be due to the learners' literacy. Auditory language is ubiquitous, whereas mastering written language need educational effort. It is possible that an illiterate person can understand auditory text with pictures, but is unable to read the corresponding written text. This will result in a strong (and absolutely trivial) modality effect.

It seems that the modality effect does not result from a unitary set of causal relationships. Instead, findings suggest that heterogeneous factors lead to similar outcomes due to rather different processing mechanisms (Schnotz, 2011; Schüler, Scheiter & Schmidt-Weigand, 2011). The ITPC model is in agreement with the split attention explanation and with the *auditory-recency* explanation of a modality effect, whereas it does not agree with an explanation via increased working memory capacity. Similar to the multimedia effect, which is counteracted by the redundancy effect as a *reversed multimedia effect*, the ITPC model can also predict a *reversed modality effect* (i.e. written text with pictures can be better for learning than spoken text with pictures), which counteracts the regular modality effect under specific conditions. Written text provides more control of cognitive processing. Readers can pause or slow down their reading, re-read difficult passages and, in this way, adapt their perceptual processing to the needs of their cognitive processing, which is much more difficult or impossible with spoken text. Thus, if a text is difficult to understand and if the accompanying picture is neither animated nor too complex and if learning time is not severely limited, the ITPC model would predict a reversed modality effect, namely better learning for pictures with

written text rather than spoken text. This is in line with recent research indicating that the modality effect occurs only under specific conditions (Gyselinck, Jamet & Dubois, 2008; Leahy, Chandler & Sweller, 2003; Stiller, Freitag, Zinnbauer & Freitag, 2009).

Interference Effects in Combining Texts and Pictures

Sequencing. Sometimes, it happens that a picture is too large and too complex to be presented simultaneously with a corresponding text. In this case, strict contiguity is hard to get. Instead, the picture has to be presented either before or after the text. Various studies have shown that it is better to present a picture before a corresponding text than after the text (Kulhavy, Stock, & Caterino, 1994). Eitel, Scheiter and Schüler (in press) have recently demonstrated with the help of eye-tracking that even a very short (i.e. less than 2 seconds) presentation of a picture can have a scaffolding function for mental model construction. The ITPC model explains this scaffolding function by the direct access of pictures to mental model construction through the mapping of analog structures in the depictive subsystem, whereas text comprehension has to make a detour through the descriptive subsystem. The sequencing effect is explained by the ITPC model through the inherent ambiguity of text. A text never describes a subject matter with enough detail to fit just one single picture or one single mental model. Instead, it allows some degrees of freedom for pictures and for mental model construction. If a mental model is constructed only from a text, the model will therefore most likely differ to some extent from a picture presented to illustrate the subject matter, even if it fully corresponds its verbal description. Thus, if the picture is presented after the text, the picture will most likely interfere with the previously text-based constructed mental model. Such interferences are avoided when the picture is presented before the text even if the learner looks only briefly at the picture to benefit from its mental model scaffolding function.

Verbal redundancy across modalities. Multimedia designers frequently try to adapt to the needs of individual learners who are assumed to prefer either spoken text or written text. They therefore present pictures simultaneously with both written text and spoken text. Learners are in this way supposed to choose their preferred sensory modality: Those who prefer to listen can focus on the spoken text, and those who prefer to read can focus on the written text. However, the ITPC model predicts that individuals do not learn better from pictures accompanied by spoken *and* by written text, but that they learn better from pictures combined with either *only spoken* or *only written* text. The model provides two reasons for this prediction. The first reason is that even if the same text is presented in an auditory manner, it is difficult for learners to ignore a simultaneously presented written text. Thus, the presentation of a picture combined with a written text will result in split of visual attention despite of the simultaneous auditory presentation of the same text. The second reason is a problem of synchronization between listening and reading. Skilled readers are often able to read a text faster than the auditory text is spoken. When they create (based on graphemic-phonemic lexical conversion) inner speech that they can hear by their inner ear, an interference between external listening and reading (i.e., internal listening) is likely to occur. Various studies of Mayer and his co-workers have demonstrated that individuals show lower performance if they learn from pictures accompanied by spoken and written text than from pictures and only spoken text (Mayer, 2009; also see Chapter 10).

Structure mapping. One and the same subject matter can often be visualized in different ways. Contrary to the dual coding theory (Paivio, 1986), which assumes that verbal and pictorial coding is generally better for learning than single coding, the ITPC model considers the form of visualization an important predictor of a multimedia effect. Pictures are only beneficial for learning if task-appropriate forms of visualization are used, whereas they are harmful in the case of task-inappropriate forms of visualization. This prediction derives

from the assumption that pictures are processed in the depictive subsystem by structure mapping. This implies that the form of visualization is mapped onto the structure of the mental model. Accordingly, the ITPC model predicts that the efficiency of a mental model for a specific task corresponds to the picture's efficiency for this task (Larkin & Simon, 1987). Corresponding empirical findings were reported by Schnotz and Bannert (2003), who studied learning from text combined with different pictures, when the pictures were informationally equivalent but used different forms of visualization. The authors found that pictures enhanced comprehension only if the learning content was visualized in a task appropriate way. If the learning content was visualized in a task-inappropriate way, the pictures interfered with the construction of a task-appropriate mental model. Thus, well-designed pictures are not only important for low prior knowledge learners who need pictorial support for mental model construction. They are also important for high prior knowledge learners, because mental model construction can be negatively affected by inappropriate forms of visualization.

Cognitive Economy

The ITPC model finally provides a framework for considerations of cognitive economy in learning from multiple external representations, especially from texts and pictures (see Chapter 20). Multiple external representations support comprehension, because each representation both constrains and elaborates the interpretation of other representations. However, understanding of each representation also creates cognitive costs. In the case of understanding multiple texts and pictures, the benefits and the costs of processing an information source depend on the ease or difficulty of using the corresponding sensory and representational channels. When more and more representations about one topic are processed, it is possible that the additional benefit for comprehension is not worth the additional cognitive costs. If the benefits from processing an additional information source are lower than the required costs, the learner will follow the principle of cognitive economy, and

he/she will not engage in further cognitive processing. Instead, the learner will consider only some representations and ignore the other ones. This could explain why individuals in self-directed learning frequently ignore information sources. This finding has been reported repeatedly in research on learning from multiple representations (Ainsworth, 1999; Sweller, van Merrienboër & Paas, 1998).

According to the ITPC model, the benefits of combining text with pictures (the multimedia effect) is not due to the superiority of dual versus single coding of information. Instead, because text is first processed in the descriptive subsystem followed by the depictive subsystem, whereas pictures are first processed in the depictive subsystem followed by the descriptive subsystem, text and pictures are assumed to have fundamentally different functions in comprehension and learning. Hochpöchler, Schnotz, Rasch, Ullrich, Horz, McElvany, Schroeder and Baumert (in press) found in an eye-tracking study of text-picture integration, that processing was primarily text driven during an initial phase of mental model construction, while brief looks at the accompanying picture indicated that pictures were only used for some scaffolding of the initial mental model. After initial mental model construction, on the contrary, the text was merely used for task-specific model updates, whereas the picture was now used intensively depending on the task at hand as an easily accessible visual tool. In other words, text processing was less task-dependent than picture processing. It seems that texts guide the reader's conceptual analysis systematically by describing the subject matter step-by-step, whereas pictures function as external cognitive tools which can be used on demand as a substitute of the subject matter.

Instructional Implications

What does the ITPC model contribute to instructional design? The model suggests various guidelines for instructional design that focus on the use of text and pictures in multimedia learning environments. Some guidelines correspond to those derived from the cognitive theory of multimedia learning (CTML) developed by (Mayer, 2009, see Chapter 3). Other guidelines go beyond the suggestions of CTML, and some further guidelines make contradicting suggestions. A fundamental commonality between the ITPC model and CTML is that both reject the idea that simple thumb rules such as the suggestion to use multiple forms of representations and multiple sensory modalities whenever possible. Instead, both views agree that instructional design for multimedia learning needs to be guided by sufficient understanding of human perception and human cognitive processing based on careful empirical research. The ITPC model suggests the following guidelines for instructional design:

- Conditional use of multimedia. Use text combined with content-related pictures, when learners have low prior knowledge, but sufficient cognitive abilities to process both the text and the pictures (cf. Chapters 7 and 24).
- Text-picture-coherence. Use pictures only when they are semantically clearly related to the content of the text (cf. Chapter 13).
- Spatial and temporal contiguity. If written text is used, present it in close spatial proximity to the picture. If spoken text is used, present it in close temporal proximity to the picture (cf. Chapter 13).
- Avoidance of redundancy. Do not combine text and pictures if learners have sufficient prior knowledge and cognitive ability to construct a mental model also from one

source of information, as the other source would be redundant for them (cf. Chapter 10).

- Text modality for animated pictures. When animations are combined with text, use spoken text instead of written text due to the fluent nature of the animation in order to avoid split of attention (cf. Chapters 9 and 22).
- Text modality for static pictures. When static pictures are used and learning time is not limited, split attention becomes less important. In this case, one should balance the advantage of auditory text (i.e. avoidance of split of attention), which predicts a positive modality effect, against the possible advantage of written text (i.e. higher control of cognitive processing), which predicts a reversed modality effect. If the text is difficult to understand, learning time is not limited, and picture complexity is low, use written text rather than spoken text (cf. Chapters 8 and 9).
- Verbal redundancy across modalities. Do not add written text that duplicates spoken text combined with pictures (cf. Chapters 13 and 20).
- Sequencing. Do not present a text that is semantically related to a picture before the picture can be observed by the learner.
- Structure-mapping. If a subject matter can be visualized by different pictures in different ways that are informationally equivalent, use a picture with the form of visualization that is most appropriate for solving future tasks.

A general message behind these suggestions is that designers of instructional material should resist the temptation to add irrelevant bells and whistles to multimedia learning environments. Simply speaking: Less can be more.

Limitations of the Integrated model and Directions for Future Research

Despite its relative complexity, the integrated model still simplifies things considerably and therefore needs further elaboration. For example, there might be multiple levels of propositional representations instead of only one level, ranging from micro-propositions (i.e. very detailed descriptions) to various levels of macro-propositions (i.e. more course-grained descriptions) based on macro-operations (van Dijk, 1980; van Dijk & Kintsch, 1983). Similarly, there might be multiple levels of mental models ranging from coarse-grained overview models to detailed partial models of high granularity. Furthermore, the interaction between the descriptive subsystem and the depictive subsystem might occur not only between propositions and a mental model as shown in Figure 4.3. When learners are highly familiar with a domain, mental models can also be constructed directly from phonological or graphemic input without a propositional detour (Perfetti & Britt, 1995). Similarly, it is possible to create a proposition directly from a perceptual representation of a visual picture without a mental model. These 'shortcuts' are not included in Figure 4.3.

Another aspect not included in the ITPC model is that that learning from text and pictures requires not only to understand the verbal and pictorial information, but also to know where which kind of information can be found. In multimedia environments, texts and pictures are frequently distributed across a complex non-linear hyperspace. In this case, the learner has to construct not only a mental model of the learning content, but also a mental model of the hyperspace.

More research is also needed to predict more precisely under which conditions the combination of text and pictures is beneficial and under which circumstances it is harmful for learning. In other words, the relative strengths of the different effects under different conditions need further specification. The effects of combining text and pictures can be

considered as a result of the different efficiency of perceptual processing and cognitive processing under specific external and internal conditions of processing. External conditions include, for example, the structure and content of the written or spoken text, text-picture coherence, text-picture redundancy, contiguity of text-picture presentation, time constraints and learning objectives. Internal conditions include, for example, prior knowledge, cognitive abilities and individual preferences. Corresponding studies should estimate the relative size of the various effects also for different types of texts and for different forms of visualization in different domains.

The ITPC model deals only with perceptual and cognitive processing of texts and instructional pictures. However, most learning material includes also decorative pictures that are expected to make the material aesthetical pleasing or perhaps to relief the learning situation (Pozzer & Roth, 2003; Takahashi, 1995). Because these pictures provide little information about the learning content, they cannot contribute much to mental model construction directly. Instead, they can be suspected to distract the learner's attention and act therefore as an impediment for learning (cf. Harp & Mayer, 1998; Sanchez & Wiley, 2006). Lenzner, Schnotz and Müller (in press) found, however, that decorative pictures captured only very little attention. However, they induced better mood, alertness and calmness with the learner, which in turn could be assumed to enhance more concentrated cognitive processing. Decorative pictures moderated the beneficial effect of instructional pictures on learning. Instructional pictures combined with decorative pictures were more beneficial for learning than those without decorative pictures. When learners had low prior knowledge, the *combined* cognitive effect of instructional pictures and affective impact of decorative pictures led to especially successful learning. Further research is needed to clarify this issue.

Future elaborations of the model should address also the learners' strategies of selecting relevant verbal or pictorial information and of giving special emphasis to specific

mental representations according to the aims of learning. As far as learners follow the principle of cognitive economy in knowledge acquisition, the efficiency of the different paths for constructing mental representations is a central concept for the analysis of strategic self-directed learning. Further research should investigate to what extent individuals follow this principle in learning from text and pictures. Individuals may prefer descriptive information processing more than depictive processing. For example, the so-called verbalizers are assumed to prefer verbal information, whereas the so-called visualizers prefer pictorial information (Kirby, Moore, & Schofield, 1988; Plass, Chun, Mayer, & Leutner, 1998). Future research should also analyze whether there are preferences with regard to the visual or the auditory modality in multimedia learning.

The ITPC model of text and picture comprehension provides a framework for the analysis of learning from multiple representations including spoken or written text, visual pictures and sound pictures. It is embedded into a broader framework of human cognition and incorporates concepts from various disciplines of cognitive science. The model aims at contributing to a deeper understanding of learning from text and pictures and to enable better-informed decisions in instructional design. Both aims require a balancing act between complexity and simplicity. The graphical representation of the ITPC model in Figure 4.3 may on the one hand be viewed as relatively complex in terms of kinds of processing and in terms of products of processing as compared with other models of comprehending texts and pictures. However, it can also be viewed as an oversimplification of the subject matter. For example, because the model deals primarily with comprehension, it only takes into account the graphemic and the phonological input lexicon, but not the phonological output lexicon (which includes the motor patterns of for producing speech sounds), although it refers to the possibility of inner speech in reading heard by the reader with his/her *inner ear*. The maxim of making things as simple as possible but not simpler (attributed to Einstein) is a special

challenge in research on multimedia learning. Future research will clarify whether the ITPC model is a useful tool for the analysis of text-picture integration.

References

Adams, B. C., Bell, L. & Perfetti, C. (1995). A trading relationship between reading skill and domain knowledge in children´s text comprehension. *Discourse Processes, 20*, 307-323.

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*, 131-152.

Atkinson, C. & Shiffrin, R.M. (1971). The control of short-term memory. *Scientific American, 225*, 82-90.

Baddeley, A. D. (1986). *Working memory.* Oxford, UK: Clarendon Press.

Baddeley, A. D. (1999). *Essentials of human memory.* Hove, UK: Psychology Press.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Science, 4,* 417-423.

Caramazza, A., Berndt, R. S., & Basili, A. G. (1983). The selective impairment of phonological processing: A case study. *Brain and Language, 18*, 128-174.

Carney, R. N. & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review, 14*, 5-26

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10,* 151-170.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204-256.

Comenius, J.A. (1999). *Orbis pictus (Facsimile of the 1887 edition).* Whitefish, MT: Kessinger Publishing.

Cooney, J. B. & Swanson, H. L. (1987) Memory and learning disabilities: An overview. In H. L. Swanson (Ed.*), Memory and learning disabilities: Advances in learning and behavioral disabilities* (pp. 1-40). Greenwich, CT: JAI.

Daneman, M. & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 9*, 561-583.

Dutke, S. (1996). Generic and generative knowledge: Memory schemata in the construction of mental models. In W. Battmann & S. Dutke (Eds.), *Processes of the molar regulation of behavior* (pp. 35-54). Lengerich: Pabst Science Publishers.

Eitel, A., Scheiter, K., & Schüler, A. (in press). The time course of information extraction from instructional diagrams. *Perceptual and Motor Skills.* Doi: 10.2466/22.23.PMS.115.6.

Ellis, A. W., & Young, A. W. (1996). *Human cognitive neuropsychology. A textbook with readings*. Hove, East Sussex: Psychology Press.

Friedman, N. P. & Miyake, A. (2000). Differential roles for visuospatial and verbal working memory in situation model construction. *Journal of Experimental Psychology: General, 129*, 61-83.

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning & Instruction, 15*, 313-331.

Graesser, A. C., Millis, K. K. & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163-189.

Gyselinck, V., Jamet, E., & Dubois, V. (2008). The role of working memory components in multimedia comprehension. *Applied Cognitive Psychology, 22,* 353-374.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90*(3), 414-434. doi:10.1037/0022-0663.90.3.414

Hochpöchler, U., Schnotz, W., Rasch, T., Ullrich, M., Horz, H., McElvany, N., Schroeder, S., & Baumert, J. (in press). Dynamics of Mental Model Construction from Text and Graphics. *European Journal of Psychology of Education*. DOI: 10.1007/s10212-012-0156-z

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge: Cambridge University Press.

Kalyuga, S., Chandler, P., & Sweller, J. (2000). Incorporating learner experience into the design of multimedia instruction. *Journal of Educational Psychology, 92,* 126-136.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394.

Kintsch, W., (1998). *Comprehension. A paradigm for cognition.* Cambridge: Cambridge University Press.

Kirby, J. R., Moore, P. J. & Schofield, N. J. (1988). Verbal and visual learning styles. *Contemporary Educational Psychology, 13,* 169-184.

Knauff, M. & Johnson-Laird, P. (2002). Visual imagery can impede reasoning. *Memory & Cognition, 30,* 363-371.

Kosslyn, S. M. (1994). *Image and brain.* Cambridge, MA: MIT Press.

Kulhavy, R.W., Stock, W.A., & Caterino, L.C. (1994). Reference maps as a framework for remembering text. In W. Schnotz & R.W. Kulhavy (Eds.), *Comprehension of graphics* (pp. 153-162). Amsterdam: Elsevier Science B.V.

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

Leahy, W., Chandler, P., & Sweller, J. (2003). When auditory presentations should and should not be a component of multimedia instruction. *Applied Cognitive Psychology, 17*, 401-418.

Lenzner, A., Schnotz, W., & Müller, A. (in press). The role of decorative pictures in learning. *Instructional Science.* DOI: 10.1007/s11251-012-9256-z.

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

Levin, J. R., Anglin, G. J. & Carney, R. N. (1987). On empirically validating functions of pictures in prose. In D. M. Willows & H. A. Houghton, (Eds.), *The psychology of illustration, Vol. 1* (pp. 51-86). New York: Springer.

Lowe, R. K. (1996). Background knowledge and the construction of a situational representation from a diagram. *European Journal of Psychology of Education, 11*, 377-397.

Mastropieri, M.A. & Scruggs, T. E. (1989). Constructing more meaningful relationships: Mnemonic instruction for special populations. *Educational Psychology Review, 1*, 83-111.

Mayer, R. E. (1997). Multimedia Learning: Are we asking the right questions? *Educational Psychologist, 32*, 1-19.

Mayer, R. E. (2001). *Multimedia learning.* New York: Cambridge University Press.

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

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology, 82*, 715-726.

Mayer, R. E., & Moreno R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90,* 312-320.

McNamara, D. S. (Ed.) (2007). *Reading comprehension strategies. Theories, interventions, and technologies.* New York: Erlbaum.

McNamara, D. S., Kintsch, E., Songer, N. B. & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1-43.

Miller, L. M. S. & Stine-Morrow, E. A. L. (1998). Aging and the effects of knowledge on on-line reading strategies. *Journal of Gerontology: Psychology Sciences, 53B,* 223-233.

Moreno, R. & Mayer, R.E. (1999). Cognitive principles of multimedia learning: the role of modality and contiguity. *Journal of Educational Psychology, 91*, 358-368.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by minimizing auditory and visual presentation modes. *Journal of Educational Psychology, 87,* 319-334.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, England: Oxford University Press.

Palmer, S. E., Rosch, E. & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and performance. Vol. 9* (p. 135-151). Hillsdale, NJ: Erlbaum.

Peirce, C.S. (1931/1958). *Collected Writings (Vols. 1-8).* (Ed. C. Hartshorne, P. Weiss & A.W Burks). Cambridge, MA: Harvard University Press.

Perfetti, C.A. & Britt, M.A. (1995). Where do propositions come from? In C.A. Weaver III, S. Mannes & C.R. Fletcher (Eds.), *Discourse comprehension. Essays in honor of Walter Kintsch* (pp. 11-34). Hillsdale, NJ: Erlbaum

Plass, J. L., Chun, D. M., Mayer, R.E. & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology, 90,* 25-36.

Pozzer, L. L., & Roth, W.-M. (2003). Prevalence, function and structure of photographs in high school biology textbooks. *Journal of Research in Science Teaching, 40*(10), 1089-1114. doi:10.1002/tea.10122

Rieben, L., & Perfetti, C. (1991). *Learning to read: Basic research and its implications.* Hillsdale, NJ: Erlbaum.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48*).* Hillsdale, NJ: Erlbaum.

Rummer, R., Schweppe, J., Fürstenberg, A., Seufert, T., & Brünken, R. (2010). Working memory interference during processing texts and pictures: Implications for the explanation of the modality effect. *Applied Cognitive Psychology, 24,* 164-176.

Sanchez, C. A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition, 34*(2), 344-355.

Schnotz, W., & Bannert, M. (1999). Einflüsse der Visualisierungsform auf die Konstruktion mentaler Modelle beim Bild- und Textverstehen [Effects of the visualization form on the construction of mental models in picture and text comprehension]. *Zeitschrift für experimentelle Psychologie, 46*, 216-235.

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

Schnotz, W. (2011). Colorful Bouquets in Multimedia Research: A Closer Look at the Modality Effect. , 269 – 276.

Schüler, A., Scheiter, K., & Schmidt-Weigand, F. (2011). Boundary conditions and constraints of the modality effect. *Zeitschrift für Pädagogische Psychologie, 25*, 211-220.

Sims, V. K. & Hegarty, M. (1997). Mental animation in the visuospatial sketchpad: Evidence from dual-tasks studies. *Memory & Cognition, 25*, 321-332.

Soederberg Miller, L. M. (2001). The effects of real-world knowledge on text processing among older adults. *Aging, Neuropsychology and Cognition, 8*, 137-148.

Stiller, K. D., Freitag, A., Zinnbauer, P., & Freitag, C. (2009). How pacing of multimedia instructions can influence modality effects: A case of superiority of visual texts. *Australasian Journal of Educational Technology, 25*, 184-203.

Sweller, J., van Merrienboër, J. G. & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychological Review, 10*, 251-296.

Takahashi, S. (1995). Aesthetic properties of pictorial perception. *Psychological Review, 102*(4), 671-683. doi:10.1037/0033-295X.102.4.671

Vallar, G. & Shallice, T. (Eds.) (1990). *Neuropsychological impairments of short-term memory*. Cambridge: Cambridge University Press.

van Dijk, T. (1980). *Macrostructures. An interdisciplinary study of global structures in discourse, interaction, and cognition*. Hillsdale, NJ: Erlbaum.

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

Van Oostendorp, H., & Goldman, S.R. (Eds.) (1999). *The construction of mental representations during reading*. Mahwah, NJ: Erlbaum.

Weaver III, C. A., Mannes, S., & Fletcher, C. R. (1995). *Discourse comprehensison. Essays in honor of Walter Kintsch*. Hillsdale, NJ: Erlbaum.

Glossary

Cognitive deep structure processing: Semantic processing of verbal and pictorial information in working memory, resulting in propositional representation and mental models.

Cognitive economy: A principle of cognitive processing that tries to reach cognitive aims with a minimum of cognitive effort.

Coherence condition: a condition for the multimedia effect, which corresponds to high semantic relatedness between text and picture.

Contiguity condition: a condition for the multimedia effect, which corresponds to high proximity of text and picture in space or in time.

Cross-modality verbal redundancy: the use of written text that duplicates spoken text combined with pictures.

Depictive representation: a form of representation that uses iconic signs (such as visual pictures) to show characteristics of a subject matter.

Descriptive representation: a form of representation that uses symbols (such as natural language) to describe characteristics of a subject matter.

Grapheme-phoneme conversion: the non-lexical conversion of letter strings into phoneme strings.

Graphemic input analysis: the identification of graphemes within visual verbal input.

Graphemic-phonemic lexical conversion: the lexicon-based conversion of whole-word letter strings into whole-word phoneme strings.

Integrated model of text and picture comprehension (ITPC model): a model of how individuals understand text and pictures presented in different sensory modalities, based on the assumption that the human perceptual system includes multiple sensory channels, whereas the cognitive system includes two representational channels: a verbal (descriptive) channel and a pictorial (depictive) channel and that these channels have limited capacity for information processing and active coherence formation.

Listening comprehension: the construction of propositional representations and mental models based on spoken text.

Mental model: a mental representation of a subject matter by an internal structure that is analogous to the subject matter.

Modality effect: students learn better from text and pictures if the text is presented as spoken rather than as written text, mainly because of avoidance of visual split attention, if specific conditions are met.

Multimedia effect: students learn better from text and pictures than from text alone, if the text is not difficult to understand, if the accompanying picture is animated or complex and if learning time is severely limited.

Parsing: syntactic-semantic analysis of spoken or written sentences (i.e. segmentation of word strings) with regard to their constituent structure.

Perceptual surface structure processing: the information transfer from the surface structure of texts and pictures to working memory, encompassing phonological or graphemic verbal information or visual or acoustic pictorial information.

Phonemic input analysis: the identification of phonemes within acoustic verbal input.

Picture-text sequencing principle: if a written text and a picture cannot be presented simultaneously, present the picture before the text instead of after the text.

Propositional representation: a mental representation of ideas expressed in a text or in a picture without reference to a specific words and phrases.

Reading comprehension: the construction of propositional representations and mental models based on written text.

Redundancy: the combination of texts and pictures when learners have sufficient prior knowledge and cognitive ability to construct a mental model also from one source only which makes the additional information source redundant for the learners and creates a reversed multimedia effect.

Reversed modality effect: students learn better from text and pictures if the text is presented as written rather than as spoken text (mainly because written text provides more control of cognitive processing than spoken text) if the text is difficult to understand and if the accompanying picture is neither animated nor too complex and if learning time is not severely limited.

Sensory register: a memory store that holds information from a specific sensory modality (e.g. the eye or the ear) for a very short time as a basis for further information processing.

Sound comprehension (auditory picture comprehension): the construction of mental models and propositional representations based on sounds (as auditory pictures).

Split attention: the use of one information channel for different sources of information.

Structure mapping: the transfer of a structure consisting of elements and relations between the elements onto another structure with different elements, but the same relations.

Text surface representation: a mental representation of a text including exact wording and syntax structure.

Visual picture comprehension: the construction of mental models and propositional representations based on visual pictures (such as drawings, maps, or graphs).

Working memory: a memory store that holds and manipulates information that is in the individual's focus of attention, including a visual store, an auditory store, a propositional store, and a spatial mental model store.

Figure Captions

Figure 4.1(a). Map of bird migration in Europe.

Figure 4.1(b). Drawing of a marsh harrier.

Figure 4.1(c). Bar graph of the marsh harrier‘s observation frequency in a Middle European habitat.

Figure 4.2. Theoretical framework for analyzing text- and picture comprehension proposed by Schnotz and Bannert (2003). A distinction is made between processing of descriptions (symbol structures) and processing of depictions (analog structures).

Figure 4.3: Integrated model of text and picture comprehension.