Frontline Learning Research
An official journal of EARLI

Frontline **L**earning **R**esearch 6 (2014) 46-66
ISSN 2295-3159

Eye Tracking Indicators of Reading Approaches in Text-Picture Comprehension

Fang Zhao[a], Wolfgang Schnotz[a], Inga Wagner[a], Robert Gaschler[a]

[a]University of Koblenz-Landau, Germany

Article received 21 March 2014 / revised 4 August 2014 / accepted 21 October 2014 / available online 27 October 2014

Abstract

Despite numerous studies on reading and multimedia comprehension, the usage of text and picture with different reading strategies has rarely become a focus of research. The current study aims to explore whether the usage of text differs from the usage of picture when readers follow different strategies of knowledge acquisition. In a within-subjects design using eye tracking, seventeen secondary school students comprehended blended text and picture materials with three different strategies. (1) Initial coherence-formation strategy, which requires students to process text and picture unguidedly. (2) Consecutive task-oriented strategy, which requires them to gather information to answer the question (which explains the task) provided after prior experience with text and picture. (3) Initial task-oriented strategy, which requires them to comprehend text and picture to solve the task equipped with the prior information about the question from the beginning. Eye tracking data showed that text and picture play different roles in these processing conditions. (1) The results are in line with the assumption that text (rather than picture) is more likely used to construct mental models in initial coherence-formation processing of text and picture. (2) Students seem to primarily rely on the picture to answer the question after the prior experience with the material with consecutive task-oriented strategy. (3) Text and picture are both used heavily when the question is presented first, enabling students to selectively process question-relevant aspects of the material at first contact.

Keywords: Multimedia Learning; Mental Model; Eye Tracking

Corresponding author: Fang Zhao, Department of Psychology, University of Koblenz-Landau, Germany, Email: zhao@uni-landau.de
Doi: http://dx.doi.org/10.14786/flr.v2i4.98

After learning to read in primary school, students are required to use their reading skills for learning from written materials. In secondary school, these materials usually include text and different kinds of pictures, such as maps and diagrams. Students therefore need skills for integrating text and picture information in order to build the required knowledge structures (Ainsworth, 1999; Kintsch & van Dijk, 1978). To support students' learning from text and pictures, we need sophisticated knowledge about the usage of text and pictures under different learning conditions. Unfortunately, there is so far not much knowledge available about integrated processing of text and pictures.

Abundant studies have explored reading strategies in text comprehension (e.g. Anderson & Pearson, 2002; Frase, 1967, 1968; Rothkopf, 1964, 1966; Rouet, 2006). The aim of reading can fundamentally influence cognitive processing due to the application of different reading strategies (Andre, 1979; Hamilton, 1985; Rickards, 1979). However, there has not been much research about reading strategies for combined comprehension of text and pictures.

The current study aims at exploring whether the usage of texts differs from the usage of pictures when readers follow different strategies of knowledge acquisition. First, we will introduce a theoretical framework of text-picture comprehension. Second, we will formulate research questions and derive hypotheses about the usage of text and pictures under the condition of different reading strategies. Third, we will describe a study that was designed to test these hypotheses. Fourth, we present the results in the light of the previously mentioned hypotheses. Fifth, we will discuss the empirical findings and analyze their relations to findings from the previous literature.

1. Theory

1.1 Theories of text-picture comprehension

Many theories focused on text-picture comprehension (TPC; Kress & Leeuwen, 1996; Rouet & Britt, 2011; Zwaan, 1998). However, there are mainly three theories specifically targeting formats of mental representations involved in TPC: (1) Dual Coding Theory (Paivio, 1986), (2) the Cognitive Theory of Multimedia Learning (Mayer, 2005) and (3) the Integrative Model of TPC (Schnotz & Bannert, 2003). The three theories share the assumption of separate channels for text and picture processing but differ on other issues. The Dual Coding Theory focuses on the referential connections between text and pictures and assumes that people can retrieve information better by using two different channels. The Cognitive Theory of Multimedia Learning assumes that people process information through an auditory-verbal channel and a visual-pictorial channel of limited capacity. Multimedia learning is assumed to include: (i) selecting relevant words; (ii) selecting relevant images; (iii) organizing the selected words into a verbal mental model; (iv) organizing the selected images into a pictorial mental model; and (v) integrating the verbal model and the pictorial model with prior knowledge into a coherent mental representation. The Integrative Model of TPC combines the concepts of multiple memory systems, multiple sensory modalities, and two kinds of representations: descriptions (such as natural language and propositional representations) and depictions (such as pictures, visual images and mental models). According to this theory, readers construct only one mental model, which contains verbal and pictorial information. Due to the importance of distinguishing descriptive and depictive representations for the analysis of reading strategies, our study is mainly inspired by the Integrative Model of TPC.

1.2 Reading Strategies of text-picture comprehension

In order to explain differences between reading strategies when learning from texts, Rickards and Denner (1978) suggested a distinction between general and specific processing. General processing deals with the global thematic coherence of the text, whereas specific processing focuses on unique information required for specific purposes. There seems to be an inherent conflict between these two kinds of processing, as Rickards and Denner found that pre-posed questions can lead to a highly selective processing, but at the expense of global understanding of the text.

According to our knowledge, this distinction between two different kinds of processing has not been applied to the combined processing of text and pictures, yet. In line with Rickards and Denner, we differentiate between *general coherence-formation processing* and *selective task-oriented processing* of text and pictures. The two kinds of processing are not meant as a strategy dichotomy. Instead, coherence-formation processing and task-oriented processing are strategy components that can be combined. As time and processing resources are limited, however, the two components cannot be both maximized at the same time. Instead, they can obtain different emphasis in the process of TPC. Thus, there is a continuum with a primarily general coherence-formation processing at the one end and a primarily selective task-oriented processing at the other end. Depending on the specific learning situation, different kinds of processing will be combined into a suitable strategy. In the following, we will consider three different learning situations:

(a) If a reader receives a text with pictures without a specific goal in mind, he/she will put the emphasis on general coherence-oriented processing. That is, he/she will try to construct coherent mental representations based on the available information. We will refer to this kind of processing as the *initial coherence-formation strategy*.

(b) If a reader has first processed a text with pictures without a specific goal in mind (i.e., applied an initial coherence-formation strategy) and then gets access to the specific questions to be answered, he/she will put emphasis on selective task-oriented processing and focus on task-relevant information. We will refer to this kind of processing as the *consecutive task-oriented strategy*.

(c) If a reader is presented specific questions before receiving a text with pictures to be used for answering these questions, he/she will put more emphasis on selective task-oriented processing. However, although processing is goal-directed from the very beginning on, some coherence-oriented processing is also required because the reader needs some understanding of what the text and the picture are about. We will refer to this kind of processing as the *initial task-oriented strategy*.

2. Research questions and hypotheses

The abovementioned strategies refer to general vs. specific information processing (Rickards & Denner, 1978), without specifying the differential roles that text vs picture might play in these strategies. In order to fill this conceptual gap, we proposed research questions and hypotheses on the usage of text vs. picture taking the described strategies into consideration. We arranged the order of presentation of (a) the question (b) the material containing text and picture, as well as potential prior exposure to the material in that way, that these context factors simulated different processing. This allowed us to compare the relative amount of text vs. picture processing among the three strategies based on eye tracking data.

As mentioned above, the initial coherence-formation strategy is comparatively general and coherence-driven. As learners have not been provided with a question to be solved based on the material, yet they might process the material in order to construct a coherent mental model covering the general content of the material. The consecutive task-oriented strategy is fairly specific and goal-driven. It is supposed to come into play when learners are provided with a question they should solve based on the material which they have already processed before. Conversely, the initial task-oriented strategy is a combination of coherence-driven and goal-driven processing. In this case, learners are provided with the question to be solved before they get access to the material. They can thus selectively process information that is relevant to the question at first contact. As there are two strategies with a higher proportion of question-oriented processing, we will focus (1) on the comparison between initial coherence-formation processing and consecutive task-oriented processing and (2) on the comparison between initial coherence-formation processing and initial task-oriented processing. Our research questions concerned potential differences between the usage of texts and the usage of pictures in the different processing. More specifically, our first research question was:

(1) Does the initial coherence-formation strategy (no question yet) differ from the consecutive task-oriented strategy (question after material) in terms of using texts and pictures?

We hypothesized that text processing differs fundamentally from picture processing with the initial coherence-formation strategy and with the consecutive task-oriented strategy. This may be routed in different functions of text and pictures in reading comprehension as well as in the effect of reading strategy. Reading through a text is a major activity to make the meaning of the content (Schmidt-Weigand et al., 2010), whereas pictures mainly serve as an external representation scaffolding the answering of questions (Eitel et al., 2013). Initial coherence-formation processing guides readers to comprehend without any task, which is more coherence-driven and general. In order to understand the main content, learners probably pay more attention to text with the initial coherence-formation processing than with the consecutive task-oriented processing. In comparison, consecutive task-oriented processing guides readers to solve the task after their prior knowledge construction, which is rather task-oriented and selective. Participants might thus pay more attention to pictures with consecutive task-oriented processing than with initial coherence-formation processing in order to solve the question.

(2) Does initial coherence-formation processing (no question yet) differ from initial task-oriented processing (question before material) in terms of using texts and pictures?

We also expected text and pictures to be comprehended differently between the initial coherence-formation strategy and the initial task-oriented strategy. The initial task-oriented strategy combines coherence formation and selective processing at first contact with the material. When readers engage in initial task-oriented processing, they, on the one hand, know the aim from the start and reading is relatively goal-directed and selective. On the other hand, they need to understand the content to search for the relevant information, which makes reading comparatively coherence-driven. In other words, the initial task-oriented strategy is more goal-driven and less coherence-driven than the initial coherence-formation strategy. As readers primarily use text for understanding, we assumed that participants focus more on text with the initial coherence-formation strategy than with the initial task-oriented strategy. As pictures can assist task solving (Larkin & Simon, 1987), we hypothesized that participants focus more on pictures and less on texts with the initial task-oriented strategy than with the initial coherence-formation strategy.

3. Method

3.1 Participants

Seventeen participants from secondary schools in Germany were included in this study (M = 13 years, SD = 3.4 years). Eleven participants were male and six were female. We used the Heller and Perleth (2000) intelligence tests on spatial ability and verbal ability. Participants were marginally above average of the norm sample in spatial ability (average T of 53.65; SD = 7.39) and verbal ability (average T of 55.59; SD = 9.08).

3.2 Materials

In a previous pilot study, we had selected 60 text-picture units from textbooks about geography and biology with 288 test questions. The units and questions were tested with 1060 students in grades 5 to 8 according to Item-Response Theory including DIF-analyses for gender, grade and school. Additionally, we carried out a rational task analysis on questions (Schnotz et al., 2011). In order to answer the questions correctly, participants need to process both text and pictures. As we adopted eye-tracking methodology, we used only a subset of the text-picture units for pragmatic reasons. Each participant received six text-picture units. The units were selected in a way that the type of image (realistic pictures vs. graphs) and the level of difficulty (easy vs. medium vs. difficult) were balanced throughout the experiment. Participants were randomly distributed to different task orders to control for sequencing effects. The selected units (see Appendix) and their average difficulties (beta-values in terms of Item-Response Theory) were as follows:

1) Banana trade: easy level (beta = -0.95) containing 95 words; realistic image in geography.

2) Legs of insects: easy level (beta = -0.75) containing 59 words; realistic image in biology.

3) Auditory: medium level (beta = 0.10) containing 122 words; graph in biology.

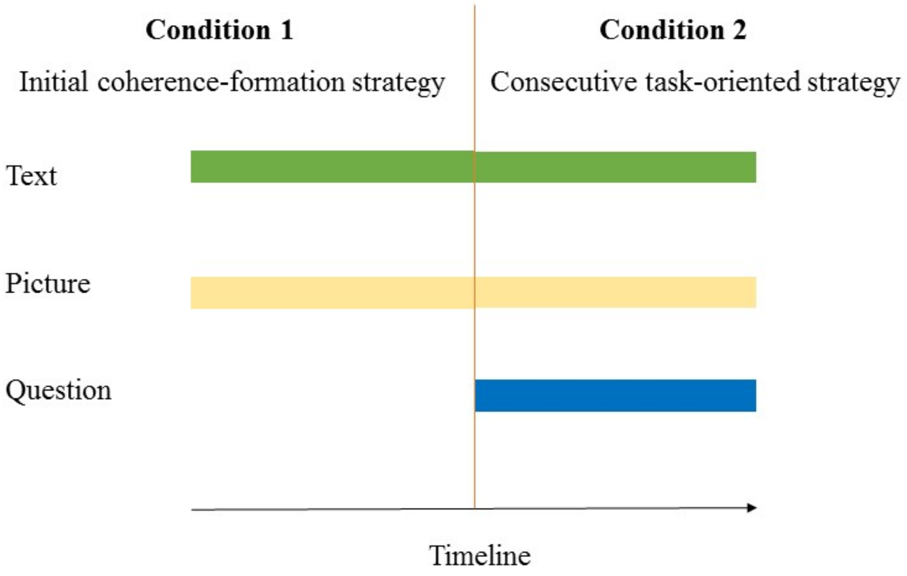
4) Pregnancy: medium level (beta = 0.39) containing 136 words; realistic image in biology.

5) Map of Europe: difficult level (beta = 1.37) containing 136 words; map in geography.

6) Savannah: difficult level (beta = 1.47) containing 170 words; graph in geography.

3.3 Stimulating specific strategies

Each participant was instructed to process six text-picture units under different conditions in order to gather eye tracking indicators of text vs. picture processing under three different processing conditions. Three units were presented without any information about the task (question) to be solved afterwards. This was expected to stimulate an initial coherence-formation strategy. After this first phase of processing, the task appeared on the screen and participants were asked to solve the task. Participants could then re-read the text and re-observe the picture under the guidance of the task. This second phase of processing was expected to stimulate a consecutive task-oriented strategy. The other three units were presented when the participants had the task to be solved already in mind. Participants read the question first. After participants had read the task, the corresponding text and pictures appeared on the screen. Therefore, participants could explore the text and the pictures under the guidance of the question from the very beginning on. This was expected to stimulate an initial task-oriented strategy.



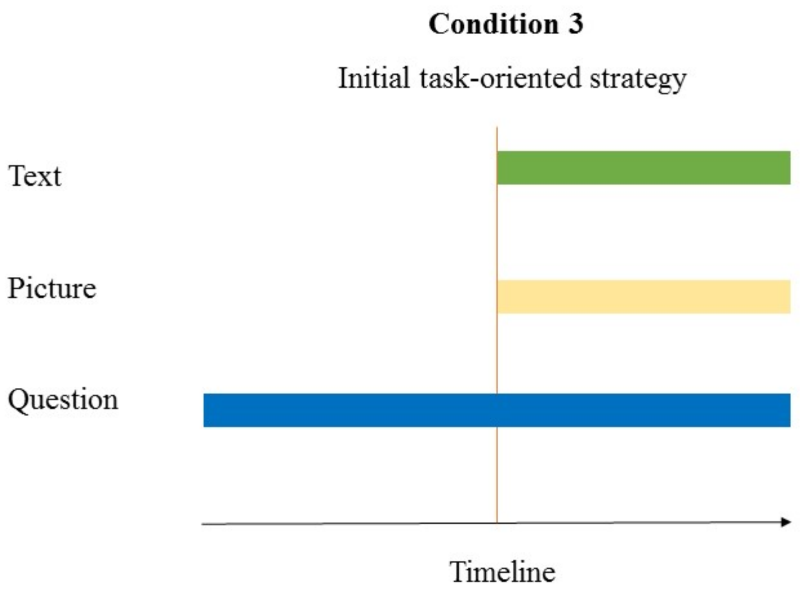


Figure 1. Design overview for the three reading conditions. The timeline indicates the time that text, picture and question appeared on the screen.

3.4 Procedure

We conducted the experiments (each taking about 45 min) individually in a lab environment with the permission of students' parents. Materials and instructions were in German (native language of the research participants). After being informed about the purpose of the study, participants took the paper-pencil IQ tests and watched an instruction video. Participants were seated at 60-65 cm distance from the 24-inch monitor of the eye tracker positioned vertically at the eye level. A 5-point calibration was conducted before participants read the material. Once the calibration was successful, the experiment would start.

The experiment included a warm-up phase and a main test. The aim of the warm-up phase was to give participants the opportunity to get familiar with the eye tracking system and with using keyboard and mouse for turning pages and answering the questions. We used a Tobii XL60 eye tracker to record eye movements at the rate of 60 Hz. The system can compensate for head movements and thus provides relatively precise data from young participants and a comfortable testing situation. For each unit, the strategy was indicated by a written instruction presented upfront. Finally, participants were thanked and rewarded with € 12 for taking part in our experiment.

3.5 Indicators of cognitive processing

Past research has established the basic assumption that cognitive processes influence and are mirrored by eye movement indicators (e.g. Gaschler, Marewski, & Frensch, in press; Godau et al., 2014). The positions that a person's eyes fixate are related to cognitive processing according to the eye-mind hypothesis and the immediacy hypothesis (Just & Carpenter, 1980). The eye-mind hypothesis assumes that eye movements can reflect information processing. The immediacy hypothesis states that the processing of information is immediate and happens directly after it is perceived. Fixation counts and fixation duration, i.e. accumulated fixation counts and the fixation time at a particular Area of Interest (AOI) are associated with the depth of cognitive processing and spatial distribution of attention (Hegarty, 1992; Rayner, 1998). Visit counts and visit duration (i.e. number of entries into and exits from an AOI and the accumulated duration of visits at an AOI) are also used in detecting reading processing as they mirror the importance of the AOI and the perceived informativeness (Jacob & Karn, 2003; Friedman & Liebelt, 1981). Time to the first fixation of an AOI (i.e. latency until an AOI is fixated for the first time) indicates readers' interest in the AOI (Goldberg

& Kotval, 1999). Transitions between AOIs (i.e., frequency of the eye movements from one AOI to the other) can mark the integration of contents presented in the AOIs (Johnson & Mayer, 2012).

3.6 Scoring

The AOIs were drawn manually using Tobii Studio software. Each task had three separated AOIs: the picture, the text and the question. The average picture AOI for the six materials covered 25.14% of the screen; and the average text AOI covered 27.72%; the average question AOI occupied 18.55%. Participants obtained one point when they answered a question correctly. They could obtain a maximum score of six points (i.e., accuracy rate of 100%) and a minimum score of zero points.

4. Results

Participants had an average of 59% accurate answers to the questions (*SD* = 25%) with initial coherence-formation strategy combined with consecutive task-oriented strategy and an average of 53% accurate answers (*SD* = 36%) with initial task-oriented strategy, $t(16) = 0.65$, $p = .52$, $d = 0.19$. The average time for comprehending text and pictures and answering the question was 1.14 seconds per word (*SD* = 0.39 second per word) from a range of 1.71 seconds per word to 0.61 seconds per word.

Most importantly, the different experimental conditions of processing strategies differed in eye fixations on text vs. picture. Considering the variation of reading speed between participants and the limited number of participants, we decided to use proportion measures for capturing the relative weight of text vs. picture (i.e. proportion of fixation counts on text, proportion of fixation time on text, proportion of visit counts on text and proportion of visit duration on text; Holmqvist et al., 2011). For example, the proportion of fixation counts on text was calculated by dividing this count by the sum of fixation counts on text and picture. Due to the dependency of text and picture, we only show the data for text with these indicators to avoid redundancy. Thus, if we refer to high proportion of fixations on text, this implies a low proportion on the picture and vice versa.

In order to explore whether the usage of text differs from the usage of picture with different strategies, we conducted two one-way repeated-measures multivariate analysis of variance (MANOVA) with four eye-tracking indicators in three reading conditions. As there were two types of task-driven strategies, two MANOVAs were performed: (1) initial coherence-formation strategy (no question yet) vs. consecutive task-oriented strategy (question after material), and (2) initial coherence-formation strategy vs. initial task-oriented strategy (question before material). The eye-tracking indicators included the proportion of fixation counts and fixation duration on text, the proportion of number of visit and visit duration on text, the time to the first fixation on text and picture and the number of transitions between text and picture.

4.1 Initial coherence-formation strategy vs. consecutive task-oriented strategy

In order to check whether the initial coherence-formation strategy (no question yet) differs from the consecutive task-oriented strategy (question after material) in fixations on text vs. picture, we performed a MANOVA and univariate analyses (ANOVAs) with the eye tracking indicators listed in Table 1. The initial coherence-formation strategy and consecutive task-oriented strategy differed significantly with respect to the relative weight on text (rather than picture) across the eye tracking indicators, $F(7, 26) = 14.10$, $p < .001$, $\eta_p^2 = .79$. As reported below, separate ANOVAs confirmed the difference between the two strategies for indicators like fixation counts and fixation duration, visit counts and visit duration and time to the first fixation.

(1) Fixation indicators during text-picture comprehension

The ANOVA revealed that the proportion of fixation counts on texts (rather than on picture) was significantly higher with initial coherence-formation processing than with consecutive task-oriented processing, $F(1, 32) = 56.51$, $p < .001$, $\eta_p^2 = .64$. Proportion of fixation counts on texts and proportion of accumulated fixation duration on texts were correlated by $r = .97$. Thus, unsurprisingly the proportion of accumulated fixation duration on text was higher with initial coherence-formation strategy than with consecutive task-oriented strategy, $F(1, 32) = 70.41$, $p < .001$, $\eta_p^2 = .69$.

(2) Visit indicators during text-picture comprehension

Participants visited the text AOI (rather than the picture AOI) more often with the initial coherence-formation strategy than with the consecutive task-oriented strategy. The proportion of number of visits on text was higher when participants engaged in initial coherence-formation processing rather than in consecutive task-oriented processing, $F(1, 32) = 7.93$, $p = .008$, $\eta_p^2 = .20$. Participants had a higher proportion of accumulated visit duration on text with initial coherence-formation processing than with consecutive task-oriented processing, $F(1, 32) = 14.24$, $p = .001$, $\eta_p^2 = .31$.

(3) Time to first fixation on text and picture

Participants fixated within a shorter time on texts with the initial coherence-formation strategy than with the consecutive task-oriented strategy, $F(1, 32) = 9.43$, $p = .004$, $\eta_p^2 = .23$. They also fixated more quickly on pictures with the initial coherence-formation strategy than with the consecutive task-oriented strategy, $F(1, 32) = 4.96$, $p = .033$, $\eta_p^2 = .13$. Pictures were fixated slightly quicker than texts: $F(1, 32) = 0.87$, $p = .358$, $\eta_p^2 = .03$ for the initial coherence-formation strategy; $F(1, 32) < 1$, for the consecutive task-oriented strategy.

(4) Transitions between text and picture

The transitions between text and picture did not show a robust difference between initial coherence-formation and consecutive task-oriented processing, $F(1, 32) = 2.34$, $p = .136$, $\eta_p^2 = .07$. With the consecutive task-oriented strategy, participants had 27% of transitions (*SD* = 10.6%) between text and picture, 25% of transitions (*SD* = 10.4%) between text and question and 48% of transitions (*SD* = 13.7%) between picture and question. In short, participants transferred their eyes slightly more often between text and picture with initial coherence-formation strategy than with consecutive task-oriented strategy. When questions were illustrated, participants mainly transferred their attention between picture and question.

Table 1

Means and standard deviations of eye tracking indicators in different goal-oriented strategies

Eye tracking indicators	Initial coherence-formation *M (SD)*	Consecutive task-oriented *M (SD)*	Initial task-oriented *M (SD)*
% Fixation counts on text	80.11 (12.81)	47.87 (12.19)	64.95 (16.4)
% Fixation time on text	81.27 (12.99)	44.00 (12.73)	66.93(15.72)
% Visit counts on text	48.29 (9.13)	39.64 (8.79)	46.16 (8.47)
% Visit time on text	66.72 (20.27)	44.88 (12.61)	69.31 (11.91)
Time to first fixation on text (sec)	2.11 (1.44)	6.60 (5.85)	2.21 (1.92)
Time to first fixation on picture (sec)	1.57 (1.92)	5.02 (6.09)	0.87 (2.07)
Average number of transitions between text and picture	21.59 (13.4)	14.94 (14.47)	28.53 (16.18)

4.2 Initial coherence-formation strategy vs. initial task-oriented strategy

Eye tracking indicators also showed that the initial coherence-formation strategy (i.e. general processing without task) differed from the initial task-oriented strategy (i.e. task was presented at the beginning) in terms of text and picture processing. The MANOVA showed the significant effect of goal-orientation on eye tracking indicators, $F(7, 26) = 3.49$, $p = .009$, $\eta_p^2 = .48$. Separate ANOVAs yielded effects of strategy condition on fixation indicators but not on visit indicators.

(1) Fixation indicators during text-picture comprehension

There was a significantly higher proportion of fixation counts on text (rather than on picture) with the initial coherence-formation strategy than with the initial task-oriented strategy, $F(1, 32) = 9.03$, $p = .005$, $\eta_p^2 = .22$. As proportion of counts and of duration of fixations were highly correlated ($r = .84$), this was mirrored by a similar effect on proportion of fixation duration on text, $F(1, 32) = 8.41$, $p = .007$, $\eta_p^2 = .21$.

(2) Visit indicators during text-picture comprehension

Participants had a similar pattern of results in visiting text and picture for both experimental strategy conditions. No difference was detected for the proportion of visit counts and visit duration on text between the initial coherence-formation strategy and the initial task-oriented strategy, $F\text{s}(1, 32) < 1$.

(3) Time to first fixation on text and picture

We did not find any difference between the experimental strategies for latency of first fixation on text or picture, $F\text{s}(1, 32) < 1.03$. Participants fixated the picture marginally sooner than the text with initial task-oriented processing, $F(1, 32) = 3.79$, $p = .06$, $\eta_p^2 = .11$.

(4) Transitions between text and picture

Transitions between text and picture did not differ when participants followed the initial coherence-formation strategy vs. the initial task-oriented strategy, $F(1, 32) = 1.48$, $p = .23$, $\eta_p^2 = .04$. For the initial task-oriented strategy, participants had 45% (SD = 13%) of transitions between text and picture, 20% (SD = 7.6%) between text and question and 35% (SD = 17.5%) between picture and question. In brief, participants transferred their eyes dominantly between text and picture, secondarily between picture and question and lastly between text and question.

5. Discussion

The current eye tracking study provides first methodological tools and results to specify the distinction between general and specific processing proposed by Rickards and Denner (1978) for processing of text vs. picture in mixed material. General processing (i.e., when the question is not known yet) deals with the global thematic coherence of the material. Pre-posed questions can lead to a highly selective processing. In order to apply this account to the processing of mixed material (text and picture), the current study examined whether text processing differs from picture processing and whether this difference is moderated by the strategies used by the learner. Specifically, we compared (1) initial coherence-formation strategy (no question yet) with consecutive task-oriented strategy (question after material) and (2) initial coherence-formation strategy with initial task-oriented strategy (question before material). A higher emphasis on text relative to picture was expected for the initial coherence-formation strategy. Pictures were assumed to lead to higher values in fixation indicators with the consecutive task-oriented processing.

We used eye tracking indicators to reveal the processing of text and picture, according to the eye-mind theory and the immediacy theory. Our results confirmed general differences of text vs. picture processing (McNamara, 2007). Importantly, eye tracking indicators of relative emphasis on text (rather than on picture) differed among the experimentally induced processing strategies. For the initial coherence-formation strategy, participants primarily fixated on text rather than on picture. As fixation count and fixation duration have been linked to the depth of cognitive processing and distribution of attention (Rayner, 1998), the result suggest the primary usage of text during mental model construction in initial coherence-formation processing. The same is true for visit counts and visit duration. As visit indicators have been linked to the importance and informativeness of the AOIs (Jacob & Karn, 2003), participants possibly consider text to be important and informative with initial coherence-formation strategy. Participants also needed little time to proceed to text and picture (i.e. time to the first fixation) with initial coherence-formation strategy. This indicator has been linked to participants' interest on text and picture when reading is general (Goldberg & Kotval, 1999). In addition, participants had frequent transitions between text and picture with initial coherence-formation strategy. According to the Integrative Model of TPC, participants may establish their mental model by integrating text and picture. This is consistent with our assumption that participants intensely processed the content to establish an initial mental model with initial coherence-formation strategy.

For the consecutive task-oriented strategy, picture was mainly used to scaffold question solving after the initial construction of the mental model. The results from fixation counts and fixation duration were consistent with the idea that pictures are primarily used when participants need to answer the question with consecutive task-oriented processing (cf. Hochpöchler et al., 2012). They had a high amount of visit counts and visit duration on pictures with consecutive task-oriented processing. It seems that participants considered pictures important and informative when they were asked to solve tasks after the initial construction of the mental model. Besides, data also revealed that participants perceived pictures sooner than texts with both processing strategies. This can be explained by pictures attracting readers' attention (Mayer, 1989; Tversky, 2001; Winn, 1989). Transition data showed that they focused their main attention on picture and question, less attention on text and picture and the least on text and question when they followed consecutive task-oriented strategy. Participants seemed to primarily use the picture to scaffold question answering. They paid less attention on text and picture because they have already constructed the initial mental model and they just need to further construct or update this mental model in order to answer the question. They paid the least attention on text and question, which implies that the text also helps participants to answer the question. Pictures possibly serve as a tool for question solving with consecutive task-oriented processing and text may be used for building and updating the mental model. These results correspond to the assumption of unequal usage of text and pictures in the Integrative Model of TPC. According to this model, pictures are mainly processed as an external representation to solve questions.

With the initial task-oriented strategy, participants invested a large amount of time on pictures because participants may have used pictures as an external tool to scaffold question answering. They visited more frequently and spent longer time on text than on picture. This might be explained by coherence-formation being supported by the text. In our design, participants need to process both text and picture to get the correct answer. Although participants were instructed with questions in initial task-oriented processing, they still needed to understand the text and the picture, thus also constructing a mental model. Therefore, text and pictures were both used with initial task-oriented processing. Also, time to the first fixation suggested that text and pictures drew participants' interest with initial task-oriented processing. Learners might integrate text and picture to build the initial mental model, as assumed in the Integrative Model of TPC. Similar patterns were shown for transitions between text and pictures with initial task-oriented strategy. Participants transferred their attention most frequently between text and picture, less between picture and question and the least between text and question with the initial task-oriented strategy. On the one hand, participants integrated text and picture with initial task-oriented strategy. On the other hand, more transitions between picture and question than between text and question support our assumption that picture is primarily used to solve the question. This result also corresponds to the assumption that text is more likely to be used for general coherence-formation processing and picture is especially used for specific task-oriented processing.

Summarizing our results, we found that text processing and picture processing differ substantially when learners are exposed to text-and-picture material. The differences are moderated by processing strategies triggered by context factors such as presentation of the question prior to vs. after the first exposure to the text-and-picture material. Likely, text is primarily used for coherence-formation processing. It assists learners to comprehend the content of the materials, which generate an initial mental model or coherent semantic representation. Picture is likely used for task-oriented specific processing. When learners have constructed the initial mental model, picture is mainly used as an external representation to update the mental model and to answer the question. When learners have the tasks beforehand, picture might serve mainly to scaffold the initial mental model construction. Future studies should provide more evidence for the link between (a) differences in fixation patterns elicited by different processing strategies and (b) the formation and usage of mental models integrating text and picture information.

In conclusion, returning to the hypotheses posed at the beginning of this study, it is now possible to state that text processing differs fundamentally from picture processing and that this difference is moderated by different reading strategies. More specifically, this study suggested that text is mainly used to build the mental model with coherence-formation general strategy. Comparatively, picture is more likely to guide readers to solve questions with task-oriented selective strategy. Similar to the previous studies on text comprehension, reading strategies also influence the comprehension of text and picture. Our findings expand the Rickards and Denner (1978) account of global vs. selective processing to the domain of mixed (picture and text) materials. The results suggest that eye tracking indicators can play a major role in assessing and scaffolding text and picture comprehension. Eye tracking indicators might be used to assess whether the learner is following an approach suitable to the current context factors (i.e. presentation of the question prior to vs. after the first exposure to the text-and-picture material). Based on such assessment, interventions guiding visual attention to areas relevant at the current processing stage can involve salient visual cues presented online during task processing (cf. Rouinfar et al. 2014).

Keypoints

- According to eye tracking indicators, text processing differs from picture processing and this difference is moderated by processing strategies.
- A high emphasis on text when processing material before the question is known, suggests that text is mainly used to build a mental model in general coherence-formation processing.
- Picture is more likely to guide readers when they need to solve a question with selective task-oriented processing.

Acknowledgments

This study is part of the on-going BiTe Project on text-picture integration, which is funded by the German Science Foundation (Grant No: SCHN 665/3-1, SCHN 665/6-1). We appreciate all the help from student participants and student assistants involved in this study. We thank for the cooperation of parents, teachers and principals. We also express our gratitude to Dr. Loredana Mihalca, Dr. Axel Zinkernagel and Dr. Thorsten Rasch for their suggestions during setting up the experiment and interpreting the data.

Appendix

Six materials displayed on Tobii eye tracker (translated from German) [each topic has three levels of questions. They were analysed on item-response theory with a one-parametric logistic model (Rasch-model): level 1 (beta = -1.15); level 2 (beta = +0.57); level 3 (beta = +1.39). We also carried out a rational task analysis, which showed the mapping procedures between text and picture: level 1 (1.25); level 2 (1.50); level 3 (3.00)].

1. Banana trade

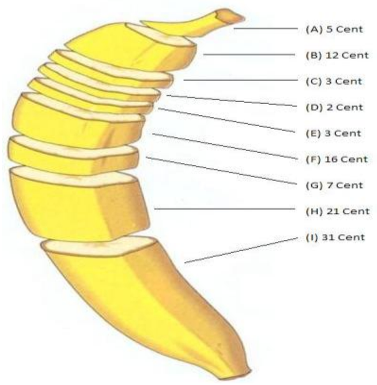


Figure A1. Banana trade.

Many people like to eat bananas. They are planted in countries like Ecuador, Costa Rica or Columbia and then exported to Europe. Undoubtedly, this is related to costs. The banana that you see in the picture costs just one euro. This euro includes...

(A) salary of the farmers;
(B) cost of the fertilizer;
(C) cost for transportation to the harbour;
(D) profit of the plantation owners;
(E) tax for bananas;
(F) cost for shipping;
(G) profit of the wholesalers;
(H) cost for storage;
(I) profit of the retailers

Question Level 1
How many cents can a retailer earn from a banana?
(3 Cent/ 21 Cent/ 31 Cent/ 7 Cent)

Question Level 2
Who do people pay the least if they buy a banana?
(profit of the wholesalers/ profit of the plantation owners/
profit of the retailers/ salary of the farmers)

Question Level 3
If we compare the farmers, retailers and wholesalers, then...
(farmers earn the most, wholesalers earn less and retailers earn the least from a banana/
retailers earn the most, wholesalers earn less and farmers earn the least from a banana/
retailers earn the most, farmers earn less and wholesalers earn the least from a banana/
wholesalers earn the most, farmers earn less and retailers earn the least from a banana)

2. Legs of insects

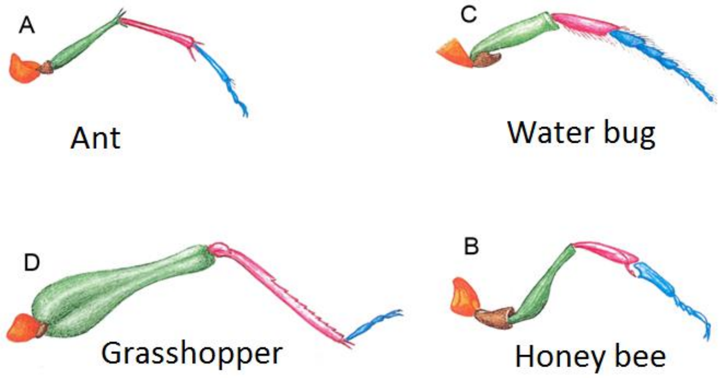


Figure A2. Legs of insects.

The legs of insects presented in figures A to D have the same structure:
Hip (orange), leg ring (brown), thigh (green), bar (pink), and foot (blue).

The legs are primarily organs for movement, which can be used for:
Running (A), swimming (B) or jumping (C). However, they can also be used for cleaning (D).

Question Level 1
Which insect has a cleaning leg?
(ant/ honey bee/ water bug/ grasshopper)

Question Level 2
Which type of leg has the longest leg ring?
(leg for cleaning/ leg for jumping/ leg for running/ leg for swimming)

Question Level 3
Does the leg for cleaning compared to the leg for jumping have…
(a longer bar but a shorter foot/ a thicker thigh but a thinner leg ring/ a longer foot but a shorter bar/ a shorter foot but a longer leg ring)

3. Auditory

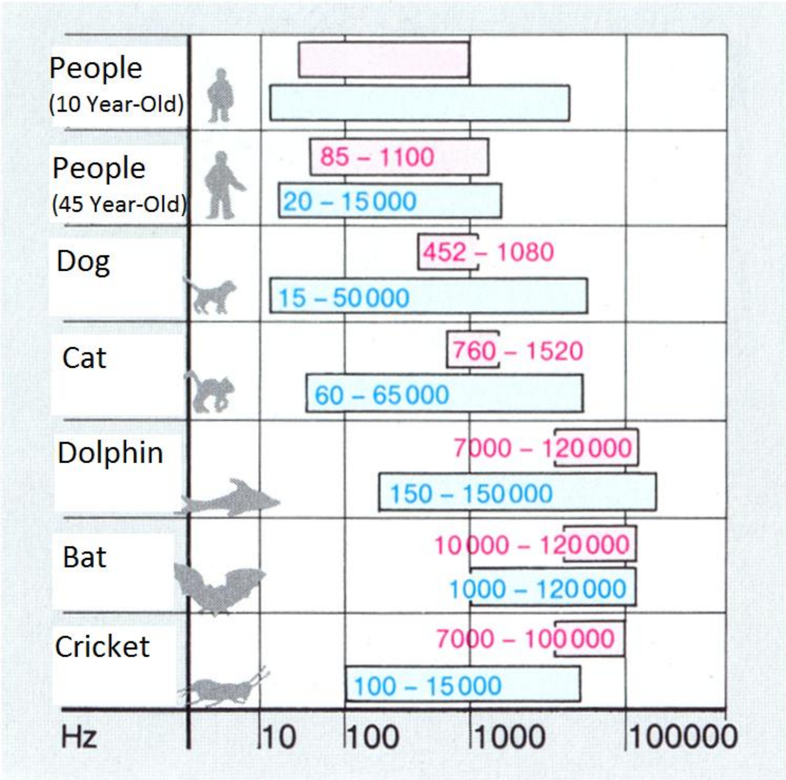


Figure A3. Auditory.

Tones and sounds are sound-waves. The faster the sound-vibrations the higher we perceive the sound/tone. The human ear can differentiate sounds/tones with low vibrations (20) and high vibrations (20 000) per second. The number of vibrations per second is called frequency; its unit is Hertz (Hz). A ten-year-old child is able to hear every sound/tone between the frequencies of 20 and 20000 Hertz. This area is called the hearing range, which is displayed in blue in the picture. Furthermore, a ten-year-old child is able to produce sounds/tones, for example by speaking, which are between 70 and 1000 Hertz. This area is called vocal range, which is displayed in pink the picture. The illustration shows the hearing and vocal ranges of different species.

Question Level 1
Which of the following species are able to perceive tones/sounds at 120 Hertz?
(dog/ cricket/ cat/ bat)

Question Level 2
Which of the following species has a vocal range for producing the lowest tone?
(human being, 45 year-old/ cat/ dog/ cricket)

Question Level 3
Which of the following four species is able to hear tones below 100 Hertz as well as produce tones below 1000 Hertz and above 1500 Hertz?
(human being, 10 year-old/ human being, 45 years old/ cricket/ cat)

4. Pregnancy

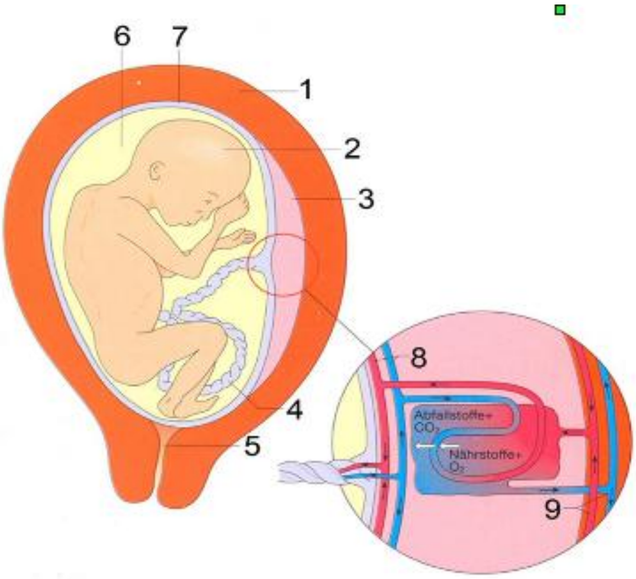


Figure A4. Pregnancy.

The child develops in the uterus, or uterine wall (1). From the fourth month of gestation it is called a fetus (2). The fetus is nourished by the placenta (3). It is here where exchange between the blood vessels of the fetus (8) and the blood vessels of the mother (9) takes place (look at zoomed area). In the blood vessels, nutrients and oxygen (O_2) as well as waste products and carbon dioxide (CO_2) are exchanged. The fetus is connected to the mother by the umbilical cord (4). The amniotic fluid protects the child. It could be called a protective-pillow for the fetus, because it helps to cushion the fetus from impact. When the mother's water membrane (7) has broken, the delivery process is initiated and the fetus will move through the cervix (5) during the labour.

Question Level 1
What is the name of the pink area?
(Placenta/ umbilical cord/ uterine wall/ amniotic fluid)

Question Level 2
Which parts do not directly link to each other?
(blood vessels of the mother and blood vessels of the child/
cervix and amniotic fluid/ umbilical cord and water membrane/ placenta and uterine wall)

Question Level 3
Which path does the blood of the fetus take after getting nutrition and oxygen (O2) from the mother? It flows …
(back through the placenta and by umbilical cord to the fetus/
back through the placenta and by the blood vessels of the fetus to the water membrane/
back through the placenta and by blood vessels of the fetus to the uterine wall/
back through the placenta and by blood vessels of the mother directly to the fetus)

5. Map of Europe

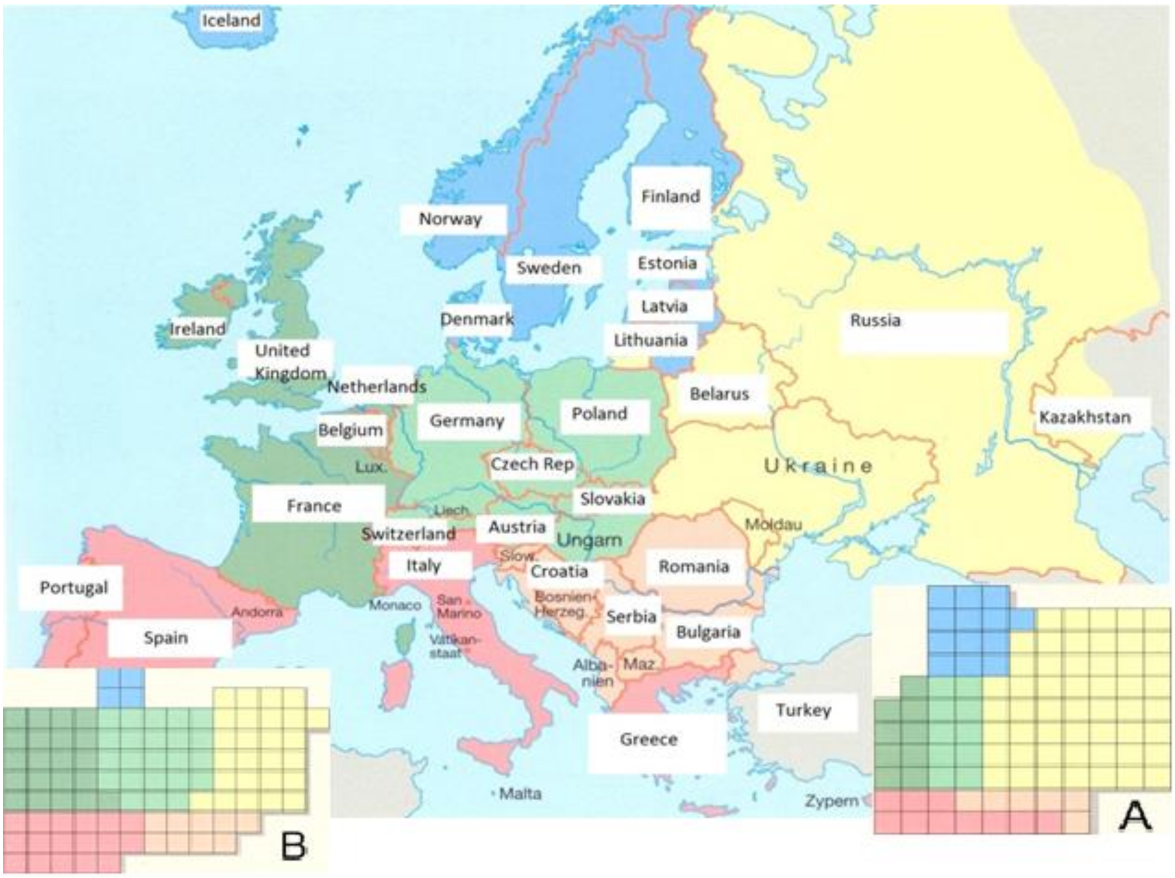


Figure A5. Map of Europe.

The big map shows the continent of Europe. Actually, Europe is not an independent continent. Together with Asia, it forms the continent "Eurasia". In the south, west and north, the border of Europe is clearly defined by the seas. The delineation in the east is more difficult because there are no natural borders. An agreement set the borders at the Ural Mountains and further south, so parts of Russia and Kazakhstan belong to both Europe and Asia.

The states of Europe are divided into different districts according to economic and geographic features. These subspaces are:
Northern Europe (blue)
Western Europe (dark green)
Central Europe (light green)
Southern Europe (red)
Eastern Europe (yellow)

In Figure A, one box represents one unit of the European area.
In Figure B, one box represents one unit of the European population.

Question Level 1
In which subspace are countries located that belong to both Europe and Asia?

(Northern Europe/ Southern Europe/ Middle Europe/ Eastern Europe)

Question Level 2

Which subspaces have the same number of area units?
(Eastern Europe and Northern Europe/ Southern Europe and South-eastern Europe/ Middle Europe and Southern Europe/ West Europe and Middle Europe)

Question Level 3

Which subspace has one more unit of area, but four fewer units of population, when compared to Western Europe?
(Middle Europe/ Northern Europe/ Southern Europe/ South-eastern Europe)

6. Savannah

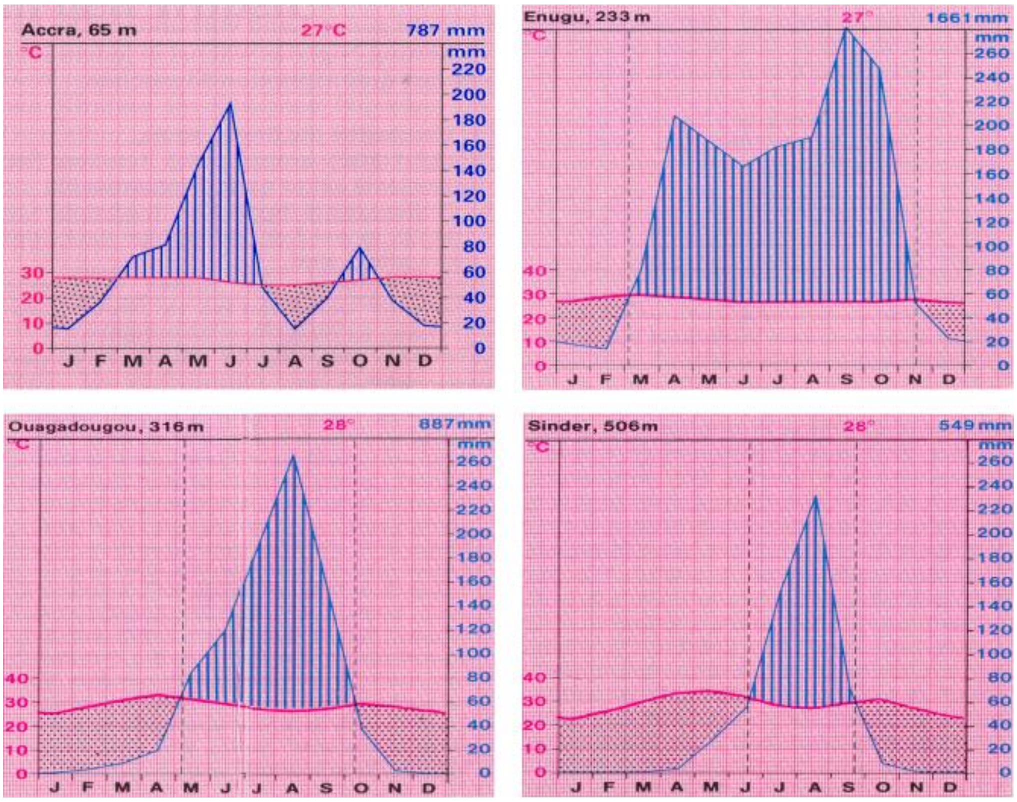


Figure A5. Savannah.

Because there are different rainy seasons, the savannah is differentiated into separate types. The city of Enugu is located in the wet savannah; Accra and Ouagadougou are located in a dry savannah and the city of Sinder in a thorn-bush savannah. Depending on the amount of rainfall (e.g. Ouagadougou 887mm rainfall per year) different food is cultivated and plants exported in each region and city. The months with enough rain for the respective plants to grow are shown in the diagrams by the areas with blue stripes above the red lines for temperature. If the line for rainfall (blue) is above the line for temperature (red), then there is more rainfall than evaporation. Each month is represented by a letter in the lower part of the diagrams, for example F = February.

The following plants need different amounts of rainfall per year for ideal growth:
Millet: 180 to 700 mm
Manioc: 500 to 2000 mm
Yams: more than 1500 mm
Peanut: 250 to 700 mm
Cotton: 700 to 1500 mm

Question Level 1
What is the amount of rainfall per year in Accra?
(887 mm/ 1661 mm/ 787 mm/ 549 mm)

Question Level 2
Which plant can grow well in Enugu?
(peanut/ yams/ cotton/ millet)

Question Level 3
Which plants will grow well in the most cities, respective of the type of savannah?
(millet/ manioc/ cotton/ peanut)

References

Anderson, R.C., & Pearson, P.D. (2002). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson, B. Rebecca, M. L. Kamil & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 255-291). Mahwah, NJ: Lawrence Erlbaum.

Ainsworth, S.E., (1999) A functional taxonomy of multiple representations. *Computers and Education*, 33(2/3), 131-152. doi:10.1016/S0360-1315(99)000299

Andre, T. (1979). Does answering higher-level questions while reading facilitate productive learning? *Review of Educational Research*, 49(2), 280-318. doi: 10.3102/00346543049002280

Eitel, A., Scheiter, K., Schüler, A., Nyström, M., & Holmqvist, K. (2013). How a picture facilitates the process of learning from text: Evidence for scaffolding. *Learning and Instruction*, 28, 48-63. doi: 10.1016/j.learninstruc.2013.05.002

Frase, L.T. (1967). Learning from prose material: Length of passage, knowledge of results, and position of questions. *Journal of Educational Psychology, 58*(5), 266-272. doi: 10.1037/h0025028

Frase, L.T. (1968). Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology*, 59(4), 244-249. doi: 10.1037/h0025947

Friedman, A., & Liebelt, L. S. (1981). On the time course of viewing pictures with a view towards remembering. In D. F. Fisher, R. A. Monty & J. Senders (Eds.), *Eye movements: cognition and visual perception* (pp. 137-155). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gaschler, R., Marewski, J. N., & Frensch, P. A. (in press). Once and for all – How people change strategy to ignore irrelevant information in visual tasks. *Quarterly Journal of Experimental Psychology*. doi: 10.1080/17470218.2014.961933

Godau, C., Wirth, M., Hansen, S., Haider, H., & Gaschler, R. (2014). From marbles to numbers – estimation influences looking patterns on arithmetic problems. *Psychology*, 5, 127-133. doi: 10.4236/psych.2014.52020

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, 24(6), 631-645. doi: 10.1016/S0169-8141(98)00068-7

Hamilton, R. J. (1985). A framework for the evaluation of the effectiveness of adjunct questions and objectives. *Review of Educational Research*, 55(1), 47-85. doi: 10.3102/00346543055001047

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(5), 1084-1102. doi: 10.1037/0278-7393.18.5.1084

Heller, K.A. & Perleth, C. (2000). *KFT 4-12+R. Kognitiver Fähigkeitstest für 4. bis 12*. Klassen, Revision. Göttingen: Beltz Test GmbH.

Hochpöchler, U., Schnotz, W., Rasch, T., Ullrich, M., Horz, H., McElvany, N., & Baumert, J. (2012). Dynamics of mental model construction from text and graphics. *European Journal of Psychology of Education*, 1(22). doi: 10.1007/s10212-012-0156-z

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Weijer, J. v. d. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford; New York: Oxford University Press.

Jacob, R. & Karn, K. S. (2003). Commentary on Section 4. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. Hyönä, R. Radach, & H. Deubel (Eds), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 573-605). Oxford: Elsevier Science.

Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied*, 18(2), 178-191. doi: 10.1037/a0026923

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychology Review*, 87, 329-354. doi: 10.1037/0033-295X.87.4.329

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363-394. doi: 10.1037/0033-295X.85.5.363

Kress, G. R., & Leeuwen, V. T. (1996). *Reading images: the grammar of visual design*. London; New York: Routledge.
Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *COGS Cognitive Science*, 11(1), 65-100. doi: 10.1111/j.1551-6708.1987.tb00863
McNamara, D. S. (Ed.). (2007). *Reading comprehension strategies: Theories, interventions, and technologies*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
Mayer, R. E. (1989). Models for Understanding. *Review of Educational Research*, 59(1), 43-64.
Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31-48). Cambridge, U.K.; New York: Cambridge University Press.
Paivio, A. (1986). *Mental representations: A dual coding approach*. New York: Oxford University Press.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422. doi: 10.1037/0033-2909.124.3.372
Rickards, J. P. (1979). Adjunct postquestions in text: A critical review of methods and processes. *Review of Educational Research*, 49(2), 181-196. doi: 10.3102/00346543049002181
Rickards, J. P., & Denner, P. R. (1978). Inserted questions as aids to reading text. *Instructional Science*, 7(3), 313-346. doi: 10.1007/BF00120936
Rothkopf, E. Z. (1964). Learning and the educational process; selected papers from the Research Conference on Learning and the Educational Process, held at Stanford University, June 22-July 31, 1964. In J. D. Krumboltz (Ed.), *Research Conference on Learning the Educational Process* (pp. 193-221). Chicago: Rand McNally.
Rothkopf, E.Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, 3(4), 241-249. doi: 10.3102/00028312003004241
Rouet, J.-F. (2006). Question answering and document search. In J.-F. Rouet (Ed.), *The skills of document use. From text comprehension to web-based learning* (pp. 93-121). Mahwah, NJ: Erlbaum.
Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano & G. Schraw (Eds.), *Text relevance and learning from text*. (pp. 19-52). Charlotte, NC, US: IAP Information Age Publishing.
Rouinfar, A., Agra E., Larson, A. M., Rebello, N. S., & Loschky, L. C. (2014). Linking attentional processes and conceptual problem solving: visual cues facilitate the automaticity of extracting relevant information from diagrams. *Frontiers in Psychology*, *5*:1094. doi:10.3389/fpsyg.2014.01094
Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). Explaining the modality and contiguity effects: New insights from investigating students' viewing behaviour. *Applied Cognitive Psychology*, 24(2), 226-237. doi: 10.1002/acp.1554
Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, 13(2003), 141-156. doi: http://dx.doi.org/10.1016/S0959-4752(02)00017-8
Schnotz, W., Ullrich, M., Hochpöchler, U., Horz, H., McElvany, N., Schröder, S., & Baumert, J. (2011). What makes text-picture integration difficult? A structural and procedural analysis of textbook requirements. *Ricerche di Psicologia*, 1, 103-135. doi: 10.1007/s10212-011-0078-1
Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 79-112). Cambridge, Mass.: MIT Press.
Winn, W. (1989). The role of graphics in training documents: Toward an explanatory theory of how they communicate. IEEE Trans. Profess. *Commun. IEEE Transactions on Professional Communication*, 32(4), 300-309. doi: 10.1109/47.44544
Zwaan, R., Radvansky, G., Hilliard, A., & Curiel, J. (1998). Constructing multidimensional situation models during reading. *Scientific Studies of Reading*, 2(3), 199-220. doi: 10.1207/s1532799xssr0203_2