Do illustrations help or harm metacomprehension accuracy?

Article *in* Learning and Instruction · December 2014

Impact Factor: 3.73 · DOI: 10.1016/j.learninstruc.2014.08.002

Allison Jaeger
Temple University

Learning and Instruction 34 (2014) 58–73

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Do illustrations help or harm metacomprehension accuracy?

Allison J. Jaeger*, Jennifer Wiley[1]

University of Illinois at Chicago, 1007 West Harrison Street M/C 285, Chicago, IL 60607, USA

ARTICLE INFO

Article history:
Received 31 December 2013
Received in revised form
29 July 2014
Accepted 10 August 2014
Available online 28 August 2014

Keywords:
Metacomprehension
Self-explanation
Multimedia

ABSTRACT

The present research examined the effect of illustrations on readers' metacomprehension accuracy for expository science text. In two experiments, students read non-illustrated texts, or the same texts illustrated with either conceptual or decorative images; were asked to judge how well they understood each text; and then took tests for each topic. Metacomprehension accuracy was computed as the intra-individual correlation between judgments and inference test performance. Results from both studies showed that the presence of decorative images can lead to poor metacomprehension accuracy. In the second study, an analysis of the cues that students reported using to make their judgments revealed that students who used comprehension-relevant cues showed more accurate metacomprehension. A self-explanation instruction did not alter either comprehension-relevant cue use or metacomprehension accuracy, although some advantages were seen when readers were prompted to self-explain from texts illustrated with conceptual images. These results suggest that students may need more explicit instruction or support to promote the use of valid cues when engaging in comprehension monitoring with illustrated text, and that seductive information such as decorative images may undermine comprehension monitoring.

1. Introduction

Accurate monitoring of comprehension (i.e. being able to differentiate between topics that have been understood well after reading from those which have not been understood well) is critical to successful self-regulated learning (Griffin, Wiley, & Salas, 2013; Thiede, Anderson, & Therriault, 2003). On any given night of homework, a student may need to read text passages about early civilizations such as the Aztecs and Incas to prepare for one test, as well as passages about photosynthesis and ecosystems to prepare for another. Effective self-regulation is especially important in these situations because it is by monitoring one's own progress while learning that decisions are made about what material needs to be restudied. If students are unable to accurately differentiate between well-learned material and less-learned material, they may waste time returning to material that is already well understood. Given the limited amount of time available for study, this may mean they will fail to restudy material that is not well understood. Despite the importance of accurate monitoring for effective self-regulated learning, students are generally poor at assessing their understanding of text passages, with typical correlations between predicted test performance and actual test performance being around .27 (Dunlosky & Lipko, 2007; Maki, 1998a; Thiede, Griffin, Wiley, & Redford, 2009). With the increasing popularity and ease of creating multimedia presentations for information in this digital age and given the widespread use of multimedia materials in educational settings, it is an important question how multimedia adjuncts may alter the metacomprehension process. Therefore, the main purpose behind this set of studies is to explore how adding illustrations to expository science texts may either improve or harm comprehension monitoring accuracy.

1.1. What is metacomprehension accuracy?

Comprehension monitoring accuracy or *metacomprehension accuracy* refers to the ability of an individual to predict how well one will do on a set of comprehension tests after reading a set of texts. Several measures of metacomprehension compare metacognitive judgments with actual performance, but each one does so in a slightly different manner. These measures include absolute accuracy, confidence bias, and relative accuracy (Maki, 1998a). Absolute accuracy is computed as the mean absolute deviation between judged and actual performance. This measure is sometimes referred to as calibration because it gives an idea of how far off a person's judgments are from actual performance. Confidence

* Corresponding author. Tel.: +1 312 339 7677; fax: +1 312 413 4122.
E-mail addresses: ajaege1@uic.edu (A.J. Jaeger), jwiley@uic.edu (J. Wiley).
[1] Tel.: +1 312 355 2501; fax: +1 312 413 4122.

http://dx.doi.org/10.1016/j.learninstruc.2014.08.002
0959-4752/

bias is a similar measure but actually concerns the direction of people's misjudgments and is sometimes referred to as over-/under confidence. This measure is computed as the signed difference between mean judgments and mean performance. Finally, relative accuracy, which is sometimes referred to as discrimination accuracy or monitoring resolution, refers to a participant's accuracy in predicting performance on one text relative to other texts (Glenberg & Epstein, 1985; Maki & Berry, 1984). As recommended by Nelson (1984), relative monitoring accuracy is computed as an intra-individual correlation between readers' judgments of learning for each text relative to the other texts, and their actual performance on each test relative to other tests. Correlations can range from −1 to +1, with correlations near 0 or below representing chance to poor accuracy. Correlations near +1 would indicate very good discrimination between texts one has understood well from those one has not. To make this concrete, imagine again a student has 4 reading assignments on a given night on the topics of Aztec Civilizations, Incan Civilizations, Photosynthesis, and Ecosystems. After having engaged in a first pass of studying, a student could be asked to rate their understanding of the 4 texts. Let's say they indicate having understood the Aztec text the best, then the Inca text, then Photosynthesis and then Ecosystems. The student would then be given test questions on each of these topics. If the test scores are aligned with the predictions (i.e. 90%, 80%, 70% and 60%) the student would be said to have perfect relative accuracy.

Although all three measures of metacomprehension accuracy are similar in that they pertain to how well a person's judgments are related to their target performance, absolute accuracy and confidence bias are statistically independent from relative accuracy (Dunlosky & Thiede, 2013). For example, a student can have good absolute accuracy or confidence bias, but poor relative accuracy. Further, absolute accuracy and confidence bias can be influenced by factors that do not affect relative accuracy measures. Specifically, absolute accuracy and confidence bias are dependent upon mean performance levels (Nelson, 1984). This can be problematic because it can in turn allow for non-metacognitive factors to influence the accuracy scores obtained, for example by things such as text or test difficulty and amount of prior knowledge (Griffin, Jee, & Wiley, 2009). Because relative accuracy is less affected by non-metacognitive factors, it is this measure that has been most commonly used in studies of metacomprehension accuracy while learning from text following the tradition established in seminal work by Glenberg and Maki (Dunlosky & Lipko, 2007; Glenberg & Epstein, 1987; Maki, 1998b; Maki & Serra, 1992; Nelson & Dunlosky, 1991; Thiede et al., 2009) and is also the measure that will be employed in the current studies.

1.2. Basic model of metacomprehension accuracy

As mentioned earlier, relative metacomprehension accuracy is partly determined by the judgments that are made by a reader. Koriat (1997) proposed the cue-utilization account to explain the accuracy of judgments-of-learning (JOLs) as a function of the cues that are used as the basis for judgments. This account posits that people have a variety of cues that they can use to predict their own test performance, and that the accuracy of these predictions hinges upon whether the chosen cues are consistent with the factors that will actually affect performance on the tests.

There is an extensive literature looking at JOLs and memory test performance for learning paired-associates such as words and their definitions and foreign language vocabulary (Metcalfe, 2002; Metcalfe & Kornell, 2003; Nelson & Dunlosky, 1991). One of the most robust findings from this literature indicates that delaying judgments serves to substantially increase the relative accuracy of JOLs compared to JOLs solicited immediately after study (see Rhodes & Tauber, 2011 for a review). Work in this area has also consistently shown that JOLs are higher for related items than unrelated items (Dunlosky & Matvey, 2001). Further, this literature has also shown support for the idea that people can make strategic study decisions based on their metacognitions (Metcalfe, 2009), a finding that demonstrates the importance of monitoring.

Although the cue-utilization account was originally formulated to explain predictions of performance in metamemory paradigms where participants are predicting their ability to recall a learned item from memory, it has also been useful in understanding the mechanisms that may be underlying meta*comprehension* accuracy, where participants are predicting whether they have learned the information that has been presented in a text. In studies that have explored metacomprehension accuracy, it has been argued the nature of "learning" that needs to be judged in the case of learning from text differs fundamentally from the previous work that used JOLs for paired-associates learning tasks. Text researchers, building on the work of Kintsch and Van Dijk (1978), have pointed out that "learning" from a text requires both memory for the text and understanding the meaning of the text, which occurs via the construction of a situation-model level representation (Rawson, Dunlosky, & Thiede, 2000; Wiley, Griffin, & Thiede, 2005). Thus, when asked to predict one's learning of a text, the task becomes more complicated than when one is asked to predict their memory performance. When asked to make JOLs when learning from text, readers have access to many cues that could affect how these judgments are made. As might be expected, one main cue they tend to use is their memory for the text (Rawson, Dunlosky, & McDonald, 2002; Thiede, Griffin, Wiley, & Anderson, 2010). However, in addition, readers also tend to rely on heuristic cues such as their interest in the topic, their prior knowledge or familiarity with the topic, or feelings of fluency while reading (Griffin et al., 2009; Rawson et al., 2000; Thiede et al., 2010) when making predictions. While these types of cues may be very salient to a reader, they are not directly related to the process of creating a mental model of the text and therefore are likely to be less valid predictors of performance on comprehension tests (Dunlosky, Rawson, & Middleton, 2005; Griffin et al., 2013; Wiley et al., 2005). The use of these cues may be responsible for the generally poor levels of metacomprehension accuracy that have been observed, around .27 (Thiede et al., 2010).

Other cues, referred to as representation-based cues (Thiede et al., 2010), develop from the process of attempting to create a mental model or situation-model-level representation of that text. These cues could include whether the person feels they could summarize the process described by the text or explain it to someone else. Although these cues are better predictors of comprehension, they tend to be used less often by students when making comprehension judgments (Thiede et al., 2010). Despite the general tendency for readers to make inaccurate judgments about comprehension, several studies have shown notable improvements in metacomprehension accuracy by putting readers in contexts designed to invoke the use of situation-model-based cues (Thiede et al., 2009). For example, readers have been shown to be more accurate when they generate keywords or summaries after a delay (Thiede & Anderson, 2003; Thiede et al., 2003). The mechanism that is suggested to underlie this phenomenon is that as time passes, surface cues decay and become less accessible, while the situation model is more robust to forgetting (Kintsch, Welsch, Schmalhofer, & Zimny, 1990). So, when keywords or summaries are generated after a delay, it helps readers to access more valid situation-model-based cues (Thiede, Dunlosky, Griffin, & Wiley, 2005). Similarly, having readers create concept maps or self-explain as part of reading has been shown to increase comprehension monitoring accuracy (Griffin, Wiley, & Thiede, 2008; Redford, Thiede, Wiley, & Griffin, 2012). Test expectancies can also affect comprehension monitoring accuracy. Thiede, Wiley, and Griffin (2011) demonstrated that metacomprehension accuracy can be improved when students are made aware that upcoming tests would ask inference-based questions. Students who expected test items about possible inferences or connections that could be drawn, instead of test items asking them to remember specific details or wording from the texts, generated more accurate predictions of comprehension. Thiede et al. (2011) argued that this increased accuracy was due to the fact that the test expectancy manipulation helped clarify the purpose for reading and therefore, once again, directed students toward more valid cues. Improved metacomprehension accuracy has also been shown among middle school students who experienced early elementary reading curricula focused on deeper comprehension processes (Thiede, Redford, Wiley, & Griffin, 2012). All of these cases suggest instances in which JOLs for learning from text become more aligned with the comprehension of information, rather than the use of memory-based or other heuristic cues, and this in turn improves metacomprehension accuracy.

1.3. *Metacomprehension accuracy for illustrated text*

All of the above studies have investigated metacomprehension accuracy and the cue-utilization account in the context of reading non-illustrated expository text passages. However, the use of plain text to convey information is becoming far from the norm in this digital age where multimedia adjuncts are now very easily and inexpensively included with prose. When students read from their textbooks or the internet, many texts are illustrated. In some domains like geology, biology and chemistry, a popular way of supporting understanding is through providing visualizations such as diagrams or schematics (e.g., Butcher, 2006). In fact, research has shown that over half of the space in middle school science textbooks is used for images or illustrations (Mayer, 1993). Although there is a fairly substantial amount of research that has explored how and when providing visualizations may affect learning from text materials (Butcher, 2006; Hegarty & Kozhevnikov, 1999; Hegarty & Sims, 1994; Mayer, 2005; Tversky, 1995, 2001), much less is known about how the presence of visualizations may affect students' metacomprehension accuracy.

From a theoretical perspective one could argue that including illustrations alongside expository text could improve metacomprehension accuracy. Levin (1981) proposed that the inclusion of images alongside text are helpful because they assist the reader in visualizing important events in the text, mentally organizing the information from the text, and in further interpreting the text. Ainsworth and Loizou (2003) showed that students who learned about the circulatory system from a combined text-and-diagram condition learned significantly more than students in a text-only condition, but that they also made more self-explanations, suggesting that illustrations were prompting the learners to engage more deeply in the construction of mental models. As mentioned earlier, previous work has shown that a self-explanation instruction can also improve metacomprehension accuracy by increasing the salience of more appropriate representation-based cues (Griffin et al., 2008). From this perspective, one could hypothesize that adding conceptually-relevant illustrations to expository text could provide an opportunity for readers' to have access to more relevant representation-based cues and therefore lead to more accurate metacomprehension.

Alternatively, one could hypothesize that including images or illustrations alongside expository text could harm metacomprehension accuracy. It has been found that adding images to texts makes reading more enjoyable (Harp & Mayer, 1998). Moreover, many educators and researchers think that adding images to make texts more interesting is important because interest and motivation are key factors that influence the selection and processing of information (Hidi, 1990). However, in many authentic educational contexts, the images that are added to illustrate texts, both in print and on the internet, are often not helpful for learning and can in fact be harmful (Lee, 2010). Linn and Hsi (2000) caution against an abundance of visual representations because they can cause confusion for students, while Mayer (1993) reported that up to 85% of science textbook illustrations lack content relevance. In fact, the common occurrence of decorative images being published alongside text to increase "visual interest" was a primary motivator for asking whether this practice might have negative effects on students. From this perspective, it could be argued that adding images to expository text could decrease metacomprehension accuracy, especially when the images are interesting or enjoyable, but not necessary for understanding the structurally important information in the text. According to the cue utilization account, images that are highly interesting, but not necessary for understanding important information from the text could be increasing readers' access to less-relevant heuristic cues. Because heuristic cues such as interest and enjoyment are not directly related to the process of creating a mental model of the text, they are not likely to be valid predictors of performance on comprehension tests and the use of these types of cues could result in less accurate metacomprehension.

In line with this perspective, Serra and Dunlosky (2010) found that students generally believe that learning from multimedia is more effective than learning from text alone, and this affects the *magnitude* of their monitoring judgments. In their Experiment 2, three groups of students were asked to read a text about lightning formation, one in which the text was paired with diagrams, one in which the text was paired with photographs of lightning, and one group in which the text was presented alone. Participants' beliefs about learning from multimedia, and initial predictions of their learning about lightning from a text that either did or did not contain images (depending on condition) were obtained prior to reading. Then, while reading the six paragraph text, participants were prompted to make a judgment of learning after each paragraph, ending with a final post-study global judgment. Finally, participants completed a memory test.

Responses to the pre-reading questionnaire indicated that all participants strongly endorsed the belief that multimedia produces better learning than single media. In addition, regardless of whether predictions were made before, during, or after reading, participants gave lower judgments in plain text conditions than either of the illustrated conditions. Importantly, test performance was only significantly better in the diagram condition; the photos did not improve test performance over the plain text condition. From these results, Serra and Dunlosky (2010) suggested that readers' beliefs about multimedia learning act as a heuristic that biases their judgments, and that the use of this *multimedia heuristic* could lead to reduced monitoring accuracy in situations where it is invalid. Although they did not report the relation of judgments to test performance in terms of absolute accuracy, and no differences were found in relative accuracy, they argued that since judgments did not differ between the two image conditions, but learning did, that students were likely using a multimedia heuristic instead of more valid processing cues as the basis for their judgments of learning. This observed effect of the mere presence of images on the judgment process would suggest that students could experience reduced monitoring accuracy when texts contain images, and this would be particularly likely to be observed in conditions where those images do not actually improve learning (Serra & Dunlosky, 2010).

Based on these ideas, the presence of illustrations, especially pertaining to less structurally relevant ideas, could be predicted to decrease relative metacomprehension accuracy because these images may give readers access to a wider range of possible cues, many of which would not be based in their situation-model-level representation of the text.

2. Experiment 1

While it has been shown that including images in expository texts can have a significant effect on the *magnitude* of metacognitive judgments (Serra & Dunlosky, 2010), their effects on *relative accuracy* measures have not yet been demonstrated. The purpose of Experiment 1 was to test whether the mere presence of two different types of images would affect the relative accuracy of participants' metacognitive judgments about learning from expository science texts. In this study, students read a set of five texts in one of three conditions: no images, conceptual images (diagrams that were directly relevant for developing a correct causal model of a scientific phenomenon or process as described by the text), or decorative images (photographs or visuals that were related to the topic of the text but did not provide explanatory information). Because the goal of this study was to look at the effect of images on relative accuracy, but not to vary the amount of information that students received across conditions, we used conceptual images that could be used to reinforce their understandings of the text, but did not offer additional information. Each student made one JOL for each text, but actually completed two types of tests in order to help assess whether the reader was relying on memory-based or comprehension-based cues in the judgment process. One test included memory-based items while the other included inference-based items. Two measures of metacognitive accuracy were computed from these ratings and tests: relative metamemory accuracy (the intra-individual correlation between ratings and test performance on memory tests), and relative metacomprehension accuracy (the intra-individual correlation between ratings and test performance on comprehension tests).

Based on previous research suggesting that conceptual diagrams may increase readers' access to more relevant representation-based cues, it was hypothesized that the inclusion of conceptual diagrams would increase relative metacomprehension accuracy (Hypothesis 1). More specifically, if readers are able to take advantage of the additional representation-based cues made available by the conceptual diagrams to diagnose their own understanding of the text, then students exposed to these images while reading should have increased relative metacomprehension accuracy as compared to students who are not exposed to conceptual images. Alternately, because seductive or decorative images may be more enjoyable and interesting for readers, but do not provide any information that can aid in mental model building, it was hypothesized that the inclusion of these types of images alongside expository text would decrease relative metacomprehension accuracy (Hypothesis 2). Again, because decorative images do not relate conceptually to learning the target content, but may be interesting or enjoyable, they may promote the use of invalid cues as a basis for judgments of learning which would lead to poorer relative metacomprehension accuracy as compared to no images or conceptual images.

2.1. Method

2.1.1. Participants

One hundred and five undergraduate students from the Introductory Psychology Subject Pool at the University of Illinois at Chicago participated in partial fulfillment of a course requirement. When participants have no variance in the judgments they make across the texts it is not possible to compute relative accuracy scores, as such three participants were dropped due to a lack of variance in their judgments resulting in a final sample of 102 participants (57 females, 45 males) with an average age of 19.43 ($SD = 2.75$). The conditions did not differ in number of science courses or composite ACT scores, $Fs < 1.4$.

2.1.2. Design

The design was a 3 (Image condition: no image, conceptual image, decorative image) × 2 (Test type: inference, memory) mixed design. Image condition was a between-subjects variable. Participants were randomly assigned to each between-subjects image condition resulting in 35 participants in the conceptual image condition, 33 participants in the no image condition, and 34 in the decorative image condition. Test type was a within-subjects variable; all students completed both types of tests.

2.1.3. Materials

2.1.3.1. Texts. The texts were five passages adapted from Thiede et al. (2011) that each described complex causal phenomena from the natural sciences (Biological evolution, Volcanoes, Ice ages, Cheesemaking, and Lightning). The texts were presented in 12 point Times New Roman font, varied in length from 800 to 1000 words and had Flesch–Kincaid grade levels of 11–12 (see Appendix for an example). A sixth text on the scientific method served as a practice text. Participants read the texts on IBM-compatible PC's in Mozilla Firefox 6.0. All browser toolbars were unavailable to the participants during the experiment. Each text was presented across three webpages so all text was visible on the screen without having to scroll. This format was chosen because previous research has indicated that the act of scrolling while reading can negatively impact learning from text, especially when readers are low in working memory capacity (Sanchez & Wiley, 2009).

2.1.3.2. Images. In the decorative and conceptual image conditions, each text was paired with only one image. Conceptual images were intended to be informationally equivalent to the text, in so far as that is possible, by providing redundant information and depicting a process involved in the scientific phenomenon described by the text. No test items required reasoning from information presented only in the diagrams. Decorative images were meant to be aesthetically pleasing and related to the topic of the text, but did not offer any information relevant for understanding the phenomenon described by the text (see the Appendix for example images). The images were always presented in the top left corner of the page and were visible throughout the reading of the entire text passage. This presentation format was the same across all five texts (see the Appendix for an example page layout).

2.1.3.3. Judgments. After reading each text, participants were instructed to make predictive judgments of learning (JOLs). The judgment specifically asked them, “If you were to take a test on the material you just read, how many questions out of 5 would you answer correctly on the test?” After responding to this question, they moved on to read the next text. Each participant made one judgment of learning for each of the five texts. This resulted in a total of 5 judgments which could range from a minimum judgment of 0 to a maximum judgment of 5.

2.1.3.4. Tests. For each text, two five-item, multiple-choice tests were created. One test consisted of memory-based items, for which answers were included explicitly in the text. An example of a memory-based item is, "How many of the world's volcanoes are located on the perimeter of the Pacific Ocean?" because the definition is found verbatim in this sentence from the text, "More than half of the world's active volcanoes above sea level encircle the Pacific Ocean to form the circum-Pacific 'Ring of Fire'." The range of difficulty for the memory-based items was 11–93 percent correct.

The other test consisted of inference-based items, which required the reader to make connections between different parts of the text to generate the answers. The design of these items was in line with suggestions from Wiley et al. (2005) about what contributes to the validity of inference items. An example of an inference-based item is, "Where is the least likely place for a volcano to occur?" The answer to this question is not explicitly stated in a single sentence, but can be inferred based on information from these two sentences from the text, "Volcanoes are not randomly distributed over the Earth's surface. Most are concentrated on the edges of continents, along island chains, or beneath the sea forming long mountain ranges." Of the twenty-five inference items, two required the reader to make an inference from a single sentence in the text by applying it to a new context, fifteen required the reader to make a connection across two to three adjacent sentences within a paragraph, three items required an inference across two sentences within the same paragraph that were not adjacent, and five items required the connection of two sentences from sequential paragraphs presented on the same page. Correct responses for two of the items were based on negations of statements in the text. The range of difficulty for the inference-based items was 14–74 percent correct.

The purpose for including two types of tests was to be able to assess both students' ability to judge the quality of their surface-level representations of the text as well as the quality of their situation-model-level representations of the texts, and to determine if adding images to text might affect whether readers are more or less likely to use memory-based or representation cues to predict their learning. This established methodological approach for correlating one set of judgments with two types of tests allows for the assessment of how students spontaneously approach making judgments-of-learning (Thiede & Anderson, 2003). The memory items tested only the surface level representation because they required recalling verbatim facts from the text, whereas the inference items relied on readers' situation model representations of the texts.

Examples of memory and inference test items are included in the Appendix. Test type was blocked, and counter-balanced so that some participants received the set of five memory tests first, and some participants received the set of five inference tests first. The topics were presented in the same order for testing as they had been seen during reading, following standard procedure in this literature (e.g., Baker & Dunlosky, 2006; Griffin et al., 2009; Thiede et al., 2005, 2009). Because the tests were all multiple choice, participants either got a 1 (correct) or 0 (incorrect) for each test item, resulting in a maximum score of 5 for each test.

2.1.4. Procedure

Prior to beginning the experiment, each participant completed an agreement to participate form. Participants completed the main portion of the experiment on the computer. The experimenter instructed each participant to click a link that allowed him or her to begin the task. This link displayed an introductory instructions page which stated,

Table 1
Mean judgment ratings, test scores, and average standard deviations by image condition for experiment 1.

	Conceptual	No image	Decorative
Average judgment ratings	2.47	2.59	2.57
Average judgment SD	.79	.77	.83
Average memory test score	2.74	2.42	2.65
Average memory test SD	1.13	1.00	1.09
Average inference test score	2.37	2.12	2.18
Average inference test SD	1.00	1.08	.98

Note. Average rating and test scores are out of a maximum of 5.

> "In this study, you will be reading a series of texts, estimating how many questions you can get correct on a five item multiple-choice test, and then taking a test to see how well you actually do. That is, you will read, predict, and test for each text."

The read-judge-test cycle used for the main portion of this experiment follows established practices for studies on metacomprehension accuracy (Dunlosky & Lipko, 2007; Thiede et al., 2009). All conditions received the same set of introductory instructions. After reading the introductory instructions all participants read the practice text, were asked to make a practice metacognitive judgment following reading, and then were given a practice inference test with five multiple choice items. In general, college students have an expectation that tests of their comprehension will rely primarily on their surface memory for the text, therefore the purpose behind providing participants with a practice inference test was to help to direct them to use the correct information when making comprehension judgments of the target texts (Thiede et al., 2011). Once participants completed the practice test they moved on to read the first target text.

After reading each text, participants made their predictive judgments. Specifically, they would read the first text then make a metacognitive judgment for that text, then they read the second text and made a metacognitive judgment; this process was continued until the participant had read and made a judgment for all 5 texts. Once all texts were read and judgments were made, students completed the two sets of multiple-choice tests. Because all of the tests were administered in the same order that the texts were read, the time between reading a text and taking the test on that text was the same across all five topics. After the tests, each participant reported their number of science courses and composite ACT scores. Finally, participants were given a debriefing sheet and thanked for their participation. The entire session took approximately 90 min to complete.

2.2. Results

2.2.1. Metacognitive judgments and test performance

Since relative monitoring accuracy is computed as a correlation between metacognitive judgments and test performance, descriptive data on these measures is reported first to ensure similar variance in both judgments and test scores across conditions following Thiede et al. (2011). As shown in Table 1, mean judgment magnitude and variance did not differ among the three image conditions, $Fs < 1$. Table 1 also reports the average test performance on memory and inference tests.[2] Importantly, participants showed

[2] A 3 (Image condition: no image, conceptual image, decorative image) × 2 (Test type: memory, inference) repeated measures ANOVA on test performance indicated a main effect for test type such that performance on the memory tests ($M = .52$) was better than performance on inference tests ($M = .45$), $F(1, 99) = 33.33$, $p < .001$. The effect for image condition did not reach significance, $F(2, 99) = 1.84$, *ns*, nor did the interaction, $F < 1$.

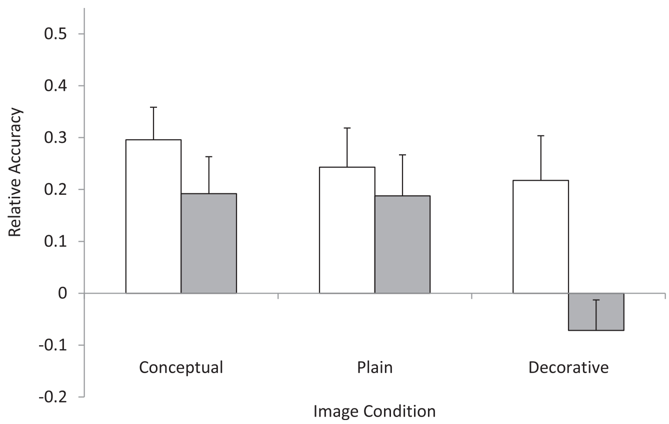


Fig. 1. Mean relative metamemory (light bars) and metacomprehension (dark bars) accuracy as a function of image condition in Experiment 1. Error bars represent the standard errors.

similar variance in their performance on the memory and inference tests across image conditions, *F*s < 1.32.

2.2.2. Relative metamemory and metacomprehension accuracy

Following the procedures of Griffin et al. (2009), two intra-individual Pearson correlations were computed for each participant: the five post-reading judgments were correlated with performance on the five memory tests (metamemory), and the five post-reading judgments were correlated with performance on the five inference tests (metacomprehension). Perfect accuracy (a correlation of 1.0) would mean that texts that were assigned a high probability of being correctly understood were actually correctly understood as evidenced by performance on the corresponding test, and texts that are assigned a low probability of being correctly understood were incorrectly understood as evidenced by performance on the corresponding text, and that the subject was able to accurately predict their relative understanding across the set of texts. Gamma correlations were also computed and showed similar patterns.[3]

As shown in Fig. 1, a 3 (Image condition: plain/no image, conceptual image, decorative image) × 2 (Accuracy measure: metamemory, metacomprehension) repeated-measures ANOVA revealed a significant main effect for accuracy measure, $F(1, 99) = 5.27$, $p < .05$, $\eta^2 = .05$. Overall, participants' judgments better predicted their memory test performance than their comprehension test performance. There was also significant main effect for image condition, $F(1, 99) = 4.01$, $p < .05$, $\eta^2 = .08$. The interaction between image condition and accuracy measure was not significant, $F(1, 99) = 1.19$, ns. Planned comparisons for each accuracy measure revealed that students' metamemory accuracy did not differ across image conditions, $F < 1$. However, metacomprehension accuracy differed across the three image conditions, $F(2, 99) = 4.63$, $p < .02$, $\eta^2 = .09$. A Tukey's honestly significant difference (HSD) test indicated that students in the decorative condition had significantly lower metacomprehension accuracy than students in the plain/no-image condition ($p < .03$) and conceptual image condition ($p < .02$). Conceptual and plain/no-image conditions did not differ.

2.3. Discussion

The results of this first experiment suggest that relative metamemory accuracy was not affected by the inclusion of images in texts, but relative metacomprehension accuracy was harmed by the presence of decorative images (Hypothesis 2). This finding is consistent with the idea that decorative images may prompt readers to rely on invalid cues for judging their own level of comprehension. Interestingly, and contrary to our hypothesis, no advantage was found in metacomprehension accuracy for readers in the conceptual image condition, suggesting that readers did not spontaneously take advantage of the conceptual illustrations as a basis for evaluating their comprehension of text (Hypothesis 1). For completeness, absolute accuracy measures were also computed and no differences were found across conditions.[4]

3. Experiment 2

Because the results from Experiment 1 indicated that participants were unable to take advantage of the conceptual images for making accurate comprehension judgments and were hurt by the presence of the decorative images, the goal of Experiment 2 was to investigate the effectiveness of an instructional manipulation meant to direct readers towards accessing more valid cues during the judgment process. Experiment 2 sought to extend and combine some aspects of the work of Ainsworth and Loizou (2003) and Griffin et al. (2008) by examining students' metacomprehension accuracy as a function of the type of image they saw while reading an expository science text and whether or not they were instructed to self-explain while reading.

Self-explanation refers to the process of generating explanations to one's self while solving problems or reading text (Chi, Deleeuw, Chiu, & Lavancher, 1994; Magliano et al., 2005). It is similar to the concept of elaboration, but with the main goal being to make sense of what one is learning rather than simply memorizing (Chi, 2000). By prompting students to make connections and note relations across sentences, to consider the meaning and relevance of each sentence, and to think about the overall purpose or theme of the text, students are more likely to understand that the purpose for reading is to construct a mental model and to make inferences (Chi et al., 1994).

As mentioned earlier, self-explanation has been shown to improve relative metacomprehension accuracy for expository text (Griffin et al., 2008). In their study, Griffin and colleagues first had all participants read through a set of five texts once. After reading all of the texts, half of the participants were instructed to self-explain while reading the set of five texts for a second time (the other condition simply re-read the texts without being instructed to self-explain). After re-reading a text the participants were prompted to judge their comprehension of that text. After reading and judging all five of the texts, participants completed a set of

[3] Using gamma correlations, a 3(Image condition: no image, conceptual image, decorative image) × 2 (Accuracy measure: metamemory, metacomprehension) repeated measures ANOVA was conducted and revealed patterns similar to when the ANOVA was conducted using Pearson correlations. There was a significant main effect for accuracy measures with metamemory ($M = .34$) being more accurate than metacomprehension ($M = .16$), $F(1, 99) = 4.52$, $p < .05$. There was also a significant main effect for image condition, $F(2, 99) = 3.63$, $p < .05$, but there was not a significant interaction, $F(2, 99) = 1.46$, *ns*. Post hoc Tukey HSD tests revealed that accuracy in the decorative condition ($M = .11$) was significantly lower than in the conceptual condition ($p = .03$) and marginally lower than the no image condition ($p = .12$), while accuracy in the conceptual ($M = .34$) and no image conditions ($M = .29$) did not differ.

[4] Although the focus of this work was on looking for differences in relative accuracy, another 3 × 2 repeated measures ANOVA was also conducted looking at participants' absolute metamemory and metacomprehension accuracy. Absolute accuracy measures were operationalized as the mean absolute deviation between a participant's metacognitive judgments and test performance across the five critical texts. Average absolute accuracy across all conditions was 1.10. No main effects or interactions were found, all *F*s < 1.70.

tests. Although this limited self-explanation instruction did not improve comprehension of the texts, participants who were instructed to self-explain were found to have more accurate metacomprehension than students who were not instructed to self-explain while reading. The authors suggested that this result was because self-explanation helped readers to focus on the construction of a mental model and in turn increased their access to cues based in their situation-model-level representations of the text. Because representation-based cues are more predictive of comprehension as assessed by an inference test, the availability and use of these cues resulted in more accurate judgments.

Based on these lines of research indicating that self-explaining while reading can help to increase readers' access and attention to situation-model-level cues, Experiment 2 sought investigate the effects of self-explanation on metacomprehension accuracy when expository texts are paired with images. All of the materials used in Experiment 2 matched those used in Experiment 1, but in addition, at the end of the study each individual was asked to describe what information they used when trying to decide if a passage should be given a high or low rating, following the procedure used in Thiede et al. (2010). These self-reports were used to provide another indicator of the cue basis that was being used for the monitoring judgments.

Based on previous research, it was predicted that self-explanation would lead to more accurate comprehension monitoring for students in all conditions because it would make situation-model-level cues more salient (Hypothesis 3). Results from Experiment 1 showed that adding decorative images to expository text resulted in less accurate comprehension monitoring, but adding a self-explanation strategy could reverse this result. Further, by adding a self-explanation instruction, students' metacomprehension accuracy in the conceptual image condition could improve to the extent that diagrams may provide an opportunity to test understanding of a text.

3.1. Method

3.1.1. Participants

One hundred and fifty undergraduate students from the Introductory Psychology Subject Pool at the University of Illinois at Chicago participated in partial fulfillment of a course requirement. Five participants were dropped due to a lack of variance in their judgments resulting in a final sample of 145 participants (88 females, 57 males) with an average age of 18.88 ($SD = 1.51$). Participants were randomly assigned to one of six conditions (plain/no explanation, plain/explanation, conceptual/no explanation, conceptual/explanation, decorative/no explanation, decorative/explanation), which did not differ in number of science courses completed or composite ACT scores, $Fs < 1$.

3.1.2. Materials and procedure

Materials and procedures were identical to those in Experiment 1 with the addition of these instructions (adapted from Griffin et al., 2008) for participants assigned to the self-explanation condition:

> In addition as you read each text, you should try to explain to yourself the meaning and relevance of each sentence or paragraph to the overall purpose of the text. Ask yourself questions like: What new information does this paragraph add? How does it relate to previous paragraphs? Does it provide important insights into the major theme of the text? Does the paragraph raise new questions in your mind? Try your best to think about these issues and ask yourself these kinds of questions about each text as you read. As you finish each paragraph, before you move on to the next paragraph, explain to yourself what that paragraph meant.

As in Griffin et al. (2008), participants were then provided a short example text with self-explanation comments after each sentence. After reading through the self-explanation instructions and example, participants proceeded through the experiment in the same manner as in Experiment 1. At the end of the study all participants were again asked to report the number of science classes they have taken and their composite ACT score, but were additionally asked to describe what information they used when trying to decide if a passage should be given a high or low rating.

3.2. Results

3.2.1. Metacognitive judgments and test performance

As shown in Table 2, mean metacognitive judgments did not differ across image or self-explanation conditions, $Fs < 1$. Importantly, similar variance in participants' judgments was seen across conditions, $Fs < 1$. Table 2 also presents average test performance on the memory and inference tests.[5] Again similar variance was seen in participants' performance on the tests across image and self-explanation conditions, $Fs < 1$.

3.2.2. Relative metamemory and metacomprehension accuracy

As in Experiment 1, the main analyses were computed using Pearson correlations. Gamma correlations were also computed and showed similar patterns.[6] A 3 (Image condition: plain/no image, conceptual image, decorative image) × 2 (Explanation condition: no explanation instruction, self-explanation instruction) × 2 (Accuracy measure: metamemory, metacomprehension) repeated-measures ANOVA revealed a significant main effect for accuracy measure, $F(1, 139) = 4.10$, $p < .05$, $\eta^2 = .03$ (see Fig. 2). Overall, participants' judgments better predicted their memory test performance than their comprehension test performance. The main effect for image condition was significant, $F(2, 139) = 3.14$, $p < .05$, $\eta^2 = .04$, but not the main effect for self-explanation condition, $F < 1$. The interaction between accuracy measure and image condition was significant, $F(2, 139) = 4.13$, $p < .02$, $\eta^2 = .06$. Planned comparisons for each accuracy measure revealed that students' metamemory accuracy did not differ across image conditions, $F < 1$. However, metacomprehension accuracy differed across the three

[5] A 3 (Image condition: conceptual image, no image, decorative image) × 2 (Test type: memory, inference) × 2 (Explanation condition: self-explanation, no self-explanation) repeated measures ANOVA on test performance indicated a main effect for test type such that performance on memory tests ($M = .53$) was better than performance on inference tests ($M = .46$), $F(1, 139) = 48.87$, $p < .001$. Neither main effect of image condition or self-explanation, nor their interaction was significant, $Fs < 1$.

[6] Analyses using gamma correlations revealed patterns similar to when the ANOVA was conducted using Pearson correlations. There was a significant main effect for accuracy measures with metamemory ($M = .35$) being more accurate than metacomprehension ($M = .19$), $F(1, 137) = 5.37$, $p < .03$. There was also a significant main effect for image condition, $F(2, 137) = 3.62$, $p < .03$, but there was no main effect for explanation condition, $F < 1$. The interaction between accuracy measure and image condition was marginal, $F(2, 137) = 2.11$, $p = .12$. Planned comparisons showed no difference in metamemory accuracy across image conditions ($F < 1$), but there was a significant difference in metacomprehension accuracy across image conditions, $F(2, 143) = 5.28$, $p < .01$. Post hoc Tukey HSD tests revealed that accuracy in the conceptual condition ($M = .40$) was significantly higher than in the decorative condition ($p < .02$) and the no image condition ($p < .02$), but accuracy in the decorative ($M = .21$) and no image ($M = .18$) conditions did not differ. There was no interaction between accuracy measure and explanation condition, $F < 1$, but the interaction between image condition and explanation condition was marginal, $F(2, 137) = 2.80$, $p = .06$. Monitoring accuracy did not differ across image conditions when participants were not instructed to self-explain, $F(2, 70) = 2.23$, *ns*, but did differ across image conditions when participants were instructed to self-explain,, $F(2, 71) = 4.27$, $p < .02$ (Conceptual: $M = .44$, Decorative: $M = .33$, Plain: $M = .09$). Finally, there was no three-way interaction, $F < 1$.

Table 2
Mean judgment ratings, test scores, average standard deviations, and proportion of participants reporting comprehension-cue use by image and explanation condition for experiment 2.

	No explanation			Explanation		
	Conceptual ($N = 24$)	Plain ($N = 25$)	Decorative ($N = 24$)	Conceptual ($N = 24$)	Plain ($N = 23$)	Decorative ($N = 25$)
Judgment ratings	2.69	2.49	2.53	2.57	2.73	2.50
Judgment SD	.72	.78	.78	.74	.74	.71
Memory test score	2.52	2.81	2.77	2.49	2.84	2.62
Memory test SD	1.20	1.08	1.08	1.12	1.10	1.06
Inference test score	2.31	2.11	2.52	2.26	2.25	2.23
Inference test SD	1.02	1.04	1.17	1.06	1.05	.94
Comprehension-cue use	.50	.40	.21	.38	.26	.32

Note. Average rating and test scores are out of a maximum of 5.

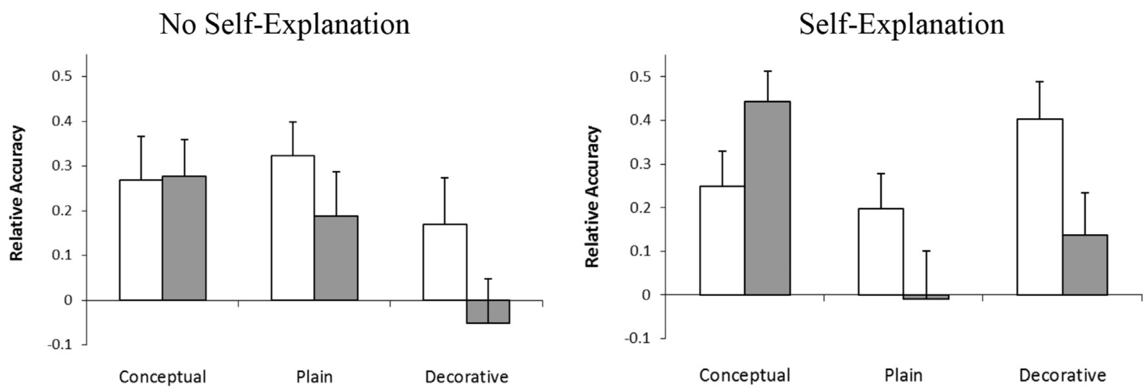


Fig. 2. Mean relative metamemory (white bars) and metacomprehension accuracy (dark bars) as a function of image and self-explanation condition in Experiment 2. Error bars represent the standard errors.

image conditions, $F(2, 142) = 5.69$, $p < .01$, $\eta^2 = .08$. Tukey's HSD indicated that students in the conceptual condition had significantly higher metacomprehension accuracy than either students in the plain/no-image or decorative image conditions, which did not differ.

In addition, the explanation condition by image condition interaction was also significant, $F(2, 139) = 4.22$, $p < .02$, $\eta^2 = .06$. Follow-up comparisons using Tukey's HSD within each explanation condition showed that monitoring was worst for the decorative image condition when readers were not prompted to self-explain (replicating Experiment 1), whereas monitoring was best for the conceptual image condition when readers were prompted to self-explain. Finally, there was no accuracy measure by self-explanation interaction and no three-way interaction, $Fs < 1$. For completeness, absolute accuracy measures were also computed and no differences were found.[7]

3.2.3. *Self-reported judgment basis*

To better understand possible reasons for the lack of improvement due to the self-explanation instruction, particularly when readers saw non-illustrated texts, the responses that students gave when asked to report what information they used when making their judgments were examined. The responses were coded into two types. If the student referred to using their ability to explain, summarize or make connections while reading the text, or referenced thinking about whether they thought they could answer questions like those they saw on the practice test as the basis for their judgments, then the student was coded as having used comprehension-based cues, following the coding from Thiede et al. (2010). If the student referred to their ability to remember the text, or their interest or prior knowledge or familiarity with the topic, then the student was coded as having used non-comprehension-based cues. Responses were independently scored by two raters who were blind to the condition. The interrater reliability that was quite high ($\kappa = .88$). Cases of disagreement were resolved through discussion. As shown in the bottom row of Table 2, self-explanation tended to lead to lower likelihood of comprehension-based cues, while conceptual images tended to lead to higher likelihood of comprehension-based cues regardless of image condition. A binary logistic regression revealed a significant effect of image condition on the likelihood of using comprehension-based cues, Wald $= 4.22$, $p < .04$. The trends for explanation condition, Wald $= 1.79$, $p < .18$, and the interaction, Wald $= 1.42$, $p < .23$, did not reach significance.

Further, cue basis did have effects on monitoring accuracy. Mean judgment and test scores for each cue basis condition are reported in Table 3. While no differences were seen in metamemory accuracy due to the use of comprehension-based cues, $t < 1$, a significant effect was found in metacomprehension accuracy, $t(143) = 6.27$, $p < .001$. To understand where cue-basis had its effects on metacomprehension accuracy, we explored whether the previously observed significant interaction between image condition and explanation condition could be seen in both cue-basis conditions. As shown in Fig. 3, there was a significant interaction between image and explanation conditions, $F(5, 89) = 3.88$, $p < .003$, $\eta^2 = .18$, for participants who reported using non-comprehension-based cues. Follow-up tests showed that the significant interaction was due to better metacomprehension accuracy in the conceptual image condition than in the plain/no-image and decorative image conditions when readers were prompted to self-explain, but no difference between image conditions when they were not. The

[7] Another 3 × 2 × 2 repeated measures ANOVA was conducted looking at participants' absolute metamemory and metacomprehension accuracy. Average absolute accuracy across all conditions was 1.15. Neither main effect for image condition or self-explanation condition, nor their interaction was significant, $Fs < 1$.

Table 3
Mean judgment ratings, test scores, and average standard deviations in experiment 2 as a function of self-explanation condition, image type, and cue basis.

	Self-explanation					
	Non-comprehension cues			Comprehension cues		
	Conceptual ($N = 15$)	Plain ($N = 17$)	Decorative ($N = 17$)	Conceptual ($N = 9$)	Plain ($N = 6$)	Decorative ($N = 8$)
Judgment ratings	2.64	2.59	2.48	2.44	3.13	2.53
Judgment SD	.74	.75	.74	.75	.70	.64
Memory test score	2.37	2.84	2.58	2.69	2.87	2.70
Memory test SD	1.10	1.13	1.07	1.15	1.00	1.05
Inference test score	2.28	2.16	2.25	2.22	2.50	2.20
Inference test SD	.95	.95	.94	1.26	1.33	.96
	No self-explanation					
	Non-comprehension cues			Comprehension cues		
	Conceptual ($N = 12$)	Plain ($N = 15$)	Decorative ($N = 19$)	Conceptual ($N = 12$)	Plain ($N = 10$)	Decorative ($N = 5$)
Judgment ratings	2.55	2.40	2.61	2.83	2.62	2.24
Judgment SD	.64	.80	.78	.80	.74	.77
Memory test score	2.47	2.77	2.82	2.57	2.86	2.56
Memory test SD	1.18	1.06	1.10	1.22	1.12	1.00
Inference test score	2.30	2.05	2.55	2.32	2.20	2.40
Inference test SD	.87	1.05	1.23	1.18	1.01	.94

Note. Average rating and test scores are out of a maximum of 5.

image by explanation condition interaction was not present when readers reported using comprehension-based cues, $F < 1$. The same analyses were conducted using gamma correlations and revealed similar patterns.[8]

3.3. Discussion

In contrast to Experiment 1, some advantages of conceptual images were seen in this study (Hypothesis 1). Readers who saw conceptual images had higher metacomprehension accuracy and were more likely to use comprehension-based cues as a basis for monitoring. The results from the no-self-explanation condition also replicated the finding from Experiment 1 that decorative images led to poor metacomprehension accuracy (Hypothesis 2). However, the self-explanation manipulation did not produce an overall benefit for monitoring accuracy (Hypothesis 3). Analyses of the self-reported judgment basis suggested that the self-explanation instruction did not lead to an increased reliance on comprehension-based cues when participants made their judgments, and if anything, tended to reduce the likelihood of using comprehension-based cues. Although the intention behind including a self-explanation instruction was to increase readers' access to comprehension-based cues and therefore increase metacomprehension accuracy, these results suggest the instruction was not effective for this purpose.

However, the cue-basis analyses revealed the more general importance of using comprehension-based cues as a basis for monitoring. When readers used comprehension-based cues, they made significantly more accurate metacomprehension judgments than when they did not. Further, this analysis allowed for the understanding of how the potential positive effects of conceptual images and potential negative effects of self-explanation were localized to conditions where readers failed to use comprehension-based cues as a basis for their judgments.

4. General discussion

Findings across both experiments show that the presence of images in texts can affect metacomprehension accuracy. Although readers did not always show an advantage from the presence of conceptual illustrations (Hypothesis 1), the presence of decorative images was shown to lead to poor metacomprehension accuracy in both experiments (Hypothesis 2). The results did not support the hypothesis that self-explanation would improve metacomprehension accuracy because it would increase access to more valid cues (Hypothesis 3).

4.1. Cue basis and self-explanation

One especially useful contribution of the current work was the focus on analyzing metacomprehension accuracy as a function of cue use, which has been suggested to be theoretically important (Koriat, 1997; Thiede et al., 2011). The results of these analyses demonstrate that participants who used comprehension-based cues had higher levels of metacomprehension accuracy. However, although it was hypothesized that self-explanation would improve metacomprehension accuracy because it would increase access to more valid cues in all conditions, the current results do not support this hypothesis (Hypothesis 3). Instructing students to self-explain did not improve metacomprehension accuracy, and did not make students any more likely to report basing their judgments on comprehension-based cues than students who were not instructed to self-explain. Further, metamemory judgments remained more accurate than metacomprehension judgments in all conditions in these studies, consistent with other work suggesting that readers' default to making JOLs for text based in their memory for the passages (Thiede et al., 2010). These results argue against the idea that either accuracy or cue-basis changed as a result of the self-explanation instruction.

At the same time, the current studies did not fully replicate the benefits that were seen for self-explanation on

[8] Analyses using gamma correlations revealed no differences seen in metamemory accuracy as a function of cue basis, $Fs < 1$. However, students who used comprehension-based cues ($M = .66$) had better metacomprehension accuracy than those who did not ($M = -.01$), $F(1, 66) = 24.27$, $p < .001$.Follow-up tests showed a significant image by explanation condition interaction ($F(5, 89) = 3.86$ $p < .01$) among readers who reported using non-comprehension-based cues. When these readers were prompted to self-explain, metacomprehension accuracy was lower in the plain ($M = -.29$) and decorative ($M = -.15$) image conditions than in the conceptual image condition ($M = .48$). No differences were seen between image conditions when these readers were not prompted to self-explain. The same interaction was not significant when readers reported using comprehension-based cues, $F < 1$.

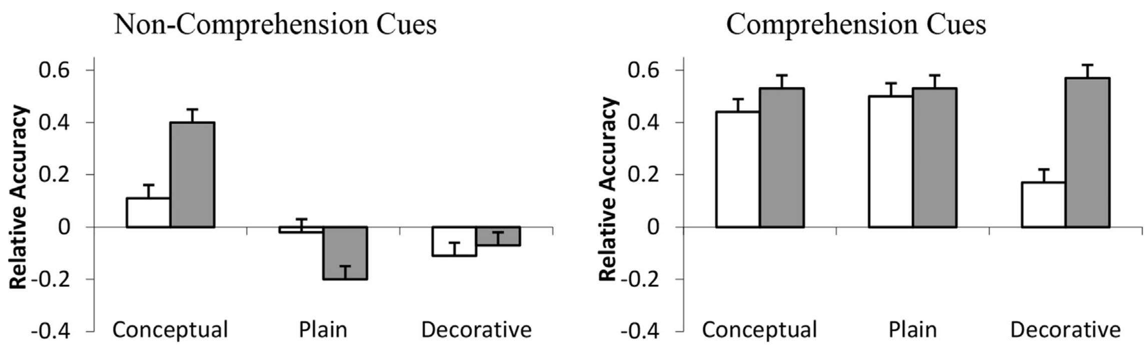


Fig. 3. Mean relative metacomprehension accuracy in the no-self-explanation (light bars) and self-explanation (dark bars) conditions in Experiment 2 as a function of image type and cue basis. Error bars represent the standard errors.

metacomprehension accuracy in Griffin et al. (2008). While the results did show that participants who were instructed to self-explain in the conceptual image condition had better relative metacomprehension accuracy, overall there was no main effect for self-explanation. One possible reason for these discrepant results could be due to differences in methods between the current study and the Griffin et al. (2008) study. Self-explanation is a highly cognitively demanding task and in Griffin et al.'s study (2008) participants were instructed to self-explain during their second reading of the texts. In the current study, participants only read the texts once and were instructed to self-explain during that initial reading. By having students re-read the text and self-explain during their second encounter with the texts, as was done in Griffin et al. (2008), they will have already developed a surface-form representation of the texts and can therefore allocate more resources to the deeper, more effortful processing required by self-explanation. The lack of benefit from self-explanation in the present results is consistent with the suggestion that the amount of cognitive resources required to read a set of texts and simultaneously self-explain may overwhelm some readers (Griffin et al., 2008). Thus, the highly cognitively demanding task of self-explaining on a first pass may be why the present study failed to fully replicate the benefit of self-explanation for metacomprehension accuracy on non-illustrated texts.

On the other hand, the effects of this minimal self-explanation instruction were in some ways similar to the results of Griffin et al. (2008), in that no main effect was seen for self-explanation on learning outcomes in either of these studies. As suggested by Griffin et al. (2008), providing readers with just an initial self-explanation instruction may not be strong enough to impact learning outcomes. In many studies on self-explanation that have found improvements in learning outcomes, students are prompted to explain aloud after each sentence is presented, and are reminded when they are not self-explaining enough, resulting in a more highly scaffolded self-explanation condition (e.g., Chi et al., 1994). In other work that has found benefits for self-explanation, students even receive training on the importance of self-explanation and what constitutes an effective self-explanation statement before working with the learning materials (e.g., Ainsworth & Burcham, 2007; McNamara, 2004). In the current study, participants were only instructed to self-explain once at the beginning of the study and were not reminded or reinforced throughout, nor were they monitored or given feedback on their explaining, which could all be reasons for the lack of benefits from the self-explanation instruction used in this study.

Given the lack of benefits seen from this minimal self-explanation instruction, an important direction for future research is figuring out what contexts actually do lead students to be more likely to use valid cues to make more accurate comprehension judgments, and whether stronger supports may be needed such as the more extensive training described above; one-on-one monitoring or feedback from a tutor; or prompting self-explanation in conjunction with other manipulations, such as instilling a test-expectancy for inference-based questions as was done in Thiede et al. (2011, 2012). In the current study, because of the online delivery setting self-explanation was implemented without one-on-one monitoring from an instructor. Under these conditions, it appears that self-explanation was fairly unhelpful. As more instruction moves online, it will become even more important to understand the critical features of self-explanation that are necessary to make it useful, and how this can be achieved or supported in online contexts.

4.2. Why does the presence of decorative images lead to poor metacomprehension?

One possible explanation for poor metacomprehension accuracy when decorative images are present is that students rely on a heuristic belief that multimedia improves comprehension (i.e. Serra & Dunlosky, 2010) rather than engaging in monitoring of their comprehension for a particular text. However, if readers simply used the presence of illustrations to inflate their monitoring judgments, then poorer absolute accuracy should have been seen. Instead, the presence of images may have lead to more idiosyncratic judgments or the use of more invalid cues, which could have been exacerbated by the self-explanation instruction particularly among readers who did not rely on comprehension-based cues. When these readers were instructed to engage in self-explanation, better relative metacomprehension accuracy was seen only in the conceptual image condition.

Another possible explanation for poor metacomprehension accuracy when decorative images are present is that students may have been “seduced” into relying too much on the images as a source for their comprehension judgments. Although images or illustrations may be included alongside text with the best of intentions, textbooks often present students with a variety of images that present a mix of important and unimportant ideas, which can be difficult for students to distinguish between (Armbruster, 1984) and this can lead to the ‘seductive details effect’ (Garner, Brown, Sanders, & Menke, 1992). The idea behind the seductive details effect is that interesting but irrelevant information competes with more relevant, structurally important information for readers' attention. Harp and Mayer (1998) showed that seductive details harm both the readers' retention of the main ideas from the text and their understanding as measured by transfer problem performance. Additionally, Hidi and Anderson (1992) showed that seductive details, which tend to be rated as highly interesting by readers, are recalled more often than the structurally important ideas from the text. Many readers experience difficulty attending to relevant information and find themselves distracted or seduced by interesting features such as decorative images (Sanchez & Wiley, 2006). As a result, the seductive appeal of decorative images may have prompted readers to rely on them when making comprehension judgments. This would have led to less accurate metacomprehension in this study because the test items focused on assessing information from the texts. The decorative images used in this study did not provide any coverage of important information from the texts and therefore judgments based on information in these images would not have been valid predictors of performance on the tests. Thus, these results are consistent with the suggestion that decorative images may be providing invalid cues, which when used as a basis for comprehension monitoring, results in poor relative accuracy.

Although the only other study that has looked at the effects of images on relative metacomprehension accuracy (Serra & Dunlosky, 2010) found no differences in accuracy due to image types, the discrepancy in results between these two lines of research can be explained by some major differences in the designs of the studies. First, Serra and Dunlosky computed relative accuracy by using judgments from multiple paragraphs of a single 500-word text about lighting formation, whereas the current work computed relative accuracy across five texts on different scientific topics that ranged from 800 to 1000 words each. This is an important distinction to note for several reasons. First, the pieces of text that students are judging their comprehension for in Serra and Dunlosky's work are much shorter and have fewer major concepts in each of them than the longer texts used in the current study. Second, students in Serra and Dunlosky were making judgments for different parts of the same text that were all dependent on each other and on the same topic. Although some researchers have had some success with this approach (Maki, Foley, Kajer, Thompson, & Wilert, 1990; Maki, Jonas, & Kallod, 1994), other work has highlighted the difficulties with attempting to assess relative accuracy using sub-sections within a single text (Thiede et al., 2009; Wiley et al., 2005). Any of these issues could have limited the ability to find relative accuracy differences in the Serra and Dunlosky study as compared to the current studies.

Another major difference between Serra and Dunlosky (2010) and the current work is the type of test that was used to assess learning. Serra and Dunlosky employed test items that could be answered with information taken verbatim from the different paragraphs of the single text. These items would be similar to the memory items used in the current studies. However, Serra and Dunlosky did not assess performance on items that required connections across different parts of the text which would have been similar to the inference items used in the current studies. In fact, when relative accuracy for the memory tests (metamemory) was assessed in the current work, no differences were found across the image conditions, essentially replicating the results of Serra and Dunlosky. However, where differences did appear in the current work was when looking at relative accuracy for the inference tests (metacomprehension), which Serra and Dunlosky were not in a position to examine.

4.3. Why didn't the presence of conceptual images lead to better metacomprehension?

While both experiments offered evidence that decorative images led to poor metacomprehension accuracy, the proposition that conceptual images could be used to support better metacomprehension accuracy was less universally supported. An overall benefit from the presence of conceptual images on metacomprehension accuracy was inconsistent across studies, and no differences were seen in test scores across image conditions in either study. Although failures to find benefits for multimedia over text alone are not atypical in the literature (Mayer, 2005), it may be the case that benefits from conceptual images on metacomprehension would be more likely to be seen under conditions where benefits are also seen on comprehension.

The failure to find advantages in either comprehension or metacomprehension accuracy when conceptual images were present could be for several reasons. One is because no test items required learners to reason about information only inferable from the diagrams. Since all important information was contained in the texts and the images provided only redundant information, there might be no reason to expect readers to do better on the test when given the conceptual images. Another possible reason could be that readers didn't attend to the images very much (i.e., they “noticed” them but didn't process them). Alternatively, readers may have processed the images but not understood them. Another class of explanations is that the images were not presented in an optimal manner. Theories of multimedia (c.f. Mayer, 2001; Schnotz, 2005) have attempted to articulate the specific conditions under which multimedia may produce greater learning than reading alone. While a full review of the principles that have been asserted goes beyond this discussion, there are several that may be germane for understanding why the conceptual image condition used here might not have been ideal. For example, according to the Split-Attention Principle (Ayres & Sweller, 2005), people learn better from multimedia when the pictures are physically and temporally integrated into the text. In the current design, the images appeared in the same place on every page, a location that was not necessarily optimal in relation to where the image would fit best within the conceptual framework of the text. As another example, it has been shown that diagrams that are not verbally integrated with the text via descriptive labels can be less effective for learning than diagrams that are clearly integrated with verbal descriptions (Mayer, Bove, Bryman, Mars, & Tapangco, 1996). Based on this research, another possible explanation for the lack of improved learning in our conceptual image condition could be related to the fact that our images were presented without verbal descriptions. This could have presented problems for learners especially if they were unable to create accurate understandings of the images on their own.

Understanding diagrams does not come easily and naturally to all people (Henderson, 1999). When learners are encouraged to self-explain from multimedia presentations, some may be disadvantaged because they may not be able to interpret the information in the images (Tversky, Zacks, Lee, & Heiser, 2000), some may have difficulty generating inferences from images (Canham & Hegarty, 2010), and some may find it challenging to integrate information across text and the illustrations (Hegarty & Just, 1993). Recent research has indicated that the allure of seductive details may also apply to graphical representations. For example, when given the opportunity to choose between representations with varying levels of detail, many people prefer representations with extraneous information despite the fact that they lead to more difficulty in comprehending and extracting information from the representations (Hegarty, Smallman, & Stull, 2012). Similarly, Lowe (1999) found that when novices were instructed to extract information from dynamic weather maps, they were more likely to extract information that attracted their attention rather than information that was relevant for building a high-quality mental model of the meteorological information. Previous research also suggests that learners experience an illusion of comprehension when learning from graphical presentations, especially those with animations (Lowe, 2003, 2004) and that students invest less effort into learning from complex graphical or dynamic materials even though these kinds of materials are more cognitively demanding (Lewalter, 2003). In line with this idea, Salomon (1984) showed that children feel more efficacious with television as compared to print, and report print being more difficult than television. However, children in the print condition showed better inference-making than children in the television condition. The present results extend this previous work by showing that readers may not be able to differentiate between their feelings of efficacy and their actual level of understanding and effort, which could be leading them to make inaccurate judgments about their comprehension when images are present.

Although the conceptual diagrams that were presented did not lead to higher test scores, they still could have been useful for making self-assessments of learning. Specifically, by providing readers with a second representation of the important information, they could have provided readers with something to compare and test their own mental representation against. One key idea behind accurate monitoring is that it allows a reader to be more aware of what they understand and do not understand; having accurate monitoring does not require that a reader already has full comprehension. From this perspective, if readers were using the diagrams to test the mental models they developed from the text, they should have been more likely to see flaws in their own representations and therefore make comprehension judgments that were better aligned with their test performance. An important goal for future research is to find the conditions under which the presence of conceptual images can be leveraged to lead to more accurate monitoring of understanding while reading.

On the other hand, while it may seem obvious to assume a positive relation between instructions that promote comprehension and instructions that promote metacomprehension, there is empirical data to suggest otherwise. For example, generating keywords at a delay has been shown to improve metacomprehension accuracy, but not necessarily test performance (Thiede et al., 2003). Additionally, generating summaries immediately after reading has been shown to benefit comprehension, while only generating summaries at a delay has been shown to benefit metacomprehension. And, completing cause–effect diagrams at a delay improves metacomprehension accuracy, but not test performance (van Loon, de Bruin, van Gog, van Merriënboer, & Dunlosky, 2014). These results all suggest that conditions that may improve learning in the short term are not necessarily the same as those that may contribute to more effective self-regulation and learning in the long term (Anderson & Thiede, 2008).

Previous work has suggested that improvements in metacomprehension accuracy are most likely when readers need to actively generate cues from memory and when those cues are generated after a delay (Thiede et al., 2009). The prevailing explanation for why completing generation activities after a delay improves monitoring accuracy, is that memory for the surface and textbase representations decay quickly while situation-model level representations are more durable. So, after a delay readers are forced to rely on more valid cues as they complete the generation tasks (Thiede et al., 2005; Wiley et al., 2005). The current study attempted to improve metacomprehension accuracy immediately and online using a self-explanation instruction, with the idea being that self-explanation would increase access to situation-model cues. However, this assumption was not borne out by the results and the lack of a delay may be another possible explanation for the lack of a benefit from the presence of the conceptual diagrams. Indeed, in the Griffin et al. (2008) study where a self-explanation prompt improved metacomprehension accuracy, the improvements were seen when readers self-explained on the second pass through the texts rather than the first time though the texts (as was the case in this study). Generating an explanation from conceptual diagrams presented after a delay or during a second reading might have allowed readers to gain more benefit from their presence (c.f. van Loon et al., 2014).

4.4. Conclusions

These studies show that the presence of images in texts can affect metacomprehension accuracy. Consistent with prior work, metacomprehension accuracy was superior when readers reported using comprehension-based cues to inform their judgments, but most readers failed to spontaneously use these valid cues (Thiede et al., 2009; Wiley et al., 2005). Readers experienced poor metacomprehension accuracy in the presence of decorative images. The current results also suggested that providing readers with conceptual diagrams and/or prompting readers to self-explain was not sufficient to ensure that they would actually use that experience, or other valid cues, as a basis for their judgments of learning. Thus, the results of the present experiments offer further evidence for the need to support students toward considering valid cues when engaging in comprehension monitoring. At the same time, the research raises new questions about how students use images to gauge their understanding of text and whether students might use different categories of images in different ways. The findings provide a first step toward bridging the multimedia and instructional design literatures with research on metacomprehension. As the use of multimedia materials and online learning platforms continue to flourish in educational settings, exploring the effects of these contexts on monitoring and self-regulation will become even more important. Specifically, these findings suggest that the design elements that give online learning materials high production value may in fact be harmful to the self-regulated learning processes that students need to deploy in online contexts.

Acknowledgments

This research was funded by Grant R305B070460 from the Cognition and Student Learning Program of the Institute of Education Sciences. A special thanks to Susan Goldman, Thomas Griffin, Mike Steiff, and Keith Thiede for their comments on this project. Additional thanks to Michelle Evans, Nicole Rivera and Samantha Hicks for assistance with data collection.

Appendix. Example text, images and tests

Volcanoes

On May 18, 1980, Mount St. Helens Volcano in Washington exploded violently. As early as March 31, seismographs began recording volcanic tremor, a type of continuous, rhythmic ground shaking. Such continuous vibrations are thought to reflect subsurface movement of fluids, either gas or magma, and suggested that magma and associated gases were on the move within the volcano. Early on May 18, following a magnitude-5.1 earthquake about 1 mile beneath the volcano, the bulged, unstable north flank of Mount St. Helens suddenly began to collapse, producing the largest landslide-debris avalanche recorded. Within seconds, eruptions began. The sudden removal of the upper part of the volcano by the landslides triggered the almost instantaneous expansion (explosion) of steam and gases within the volcano. The abrupt pressure release uncorked the volcano. A strong, vertically directed explosion of ash and steam began very shortly after the lateral blast and rose very quickly. In less than 10 min, the ash column reached an altitude of more than 12 miles and began to expand into a mushroom-shaped ash cloud.

Volcanoes are not randomly distributed over the Earth's surface. Most are concentrated on the edges of continents, along island chains, or beneath the sea forming long mountain ranges. More than half of the world's active volcanoes above sea level encircle the Pacific Ocean to form the circum-Pacific "Ring of Fire." Plate tectonics tells us that the Earth's rigid outer shell is broken into a dozen or so plates. These plates are riding on currents in the hot, mobile uppermost layer of the mantle. When plates interact at their margins, important geological processes take place, such as the formation of mountain belts, volcanoes and most earthquakes.

Though hidden underwater, the global mid-ocean ridge system is the most prominent topographic feature on the surface of our planet. In 1961, scientists began to theorize that mid-ocean ridges mark structurally weak zones where ocean plates were being ripped in two. New magma from deep within the Earth rises easily through these weak zones and eventually erupts along the crest of the ridges to create new oceanic crust. This process, called seafloor spreading, has built the mid-ocean ridges. Henry Hess reasoned that the ocean basins were perpetually being "recycled," with the creation of new crust and the destruction of old oceanic lithosphere occurring simultaneously. He suggested that new oceanic crust continuously spreads away from the ridges in a conveyor belt-like motion. Many millions of years later, the oceanic crust eventually descends into the oceanic trenches – very deep, narrow canyons along the rim of the Pacific Ocean basin. The amount of crust remains constant. When a divergence of plates occurs in one area, a convergence of plates occurs in another.

There are 3 types of converging plate boundaries: Oceanic–Oceanic, Oceanic–Continental and Continental–Continental. When an oceanic–continental convergence occurs, one plate will most commonly subduct beneath the other plate creating a trench. The oceanic plate is denser than the continental plates, so the oceanic plate is usually subducted. For example, the east edge of the Juan de Fuca Plate is plunging beneath the North American Plate. As the oceanic crust is forced deep into the Earth's interior beneath the continental plate, it encounters high temperatures and pressures. The melting of the crust forms magma. Some of this newly formed magma rises toward the Earth's surface. Arcs of volcanoes usually form above a subduction zone. Earthquakes can also be caused by the collision of oceanic and continental plates. In the Philippines, the Java trench is associated with volcanic islands as well as earthquakes. Further, the movement of magma in subduction zones can also trigger deep earthquakes.

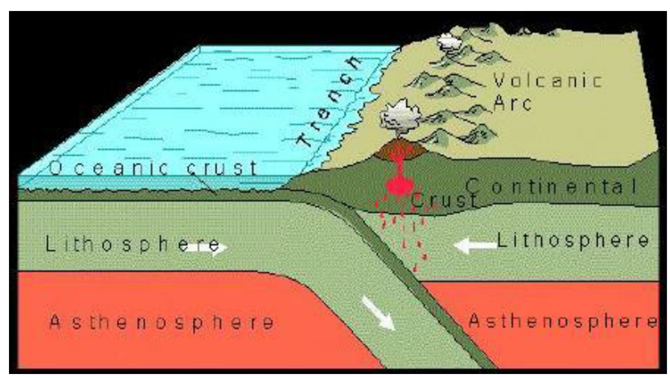
An oceanic–oceanic convergence often results in the formation of an island arc system. As one plate subducts it melts within the mantle. The magma rises to the surface of the ocean floor and forms volcanoes. If the activity continues, the volcano may grow tall enough to create an island. A continental–continental convergence generally does not involve subduction. Instead, the two plates squeeze and deform each other, resulting in a mountain range such as the Himalayas. Earthquakes are also associated with high mountain ranges where intense compression is taking place.

Scientists have defined two major types of volcanoes: shield volcanoes and stratovolcanoes. Shield volcanoes are the largest volcanoes on Earth. They are gently sloping, such as those in Hawaii. Their lavas flow great distances from the active vents. Hawaiian magmas have a low viscosity, and gases can escape prior to an eruption. Like most oceanic volcanoes, their magma comes from the melting of crust in the ocean plates. Hawaiian eruptions are noted for their non-explosive nature and approachability.

Stratovolcanoes are typically located near convergent plate boundaries where subduction is occurring, particularly around the Pacific basin. The magma produced by subduction is generally high in viscosity. The high viscosity does not allow gas to readily escape from the magma. When the magma reaches the vent of the volcano, gas bubbles begin to form and to grow. The rapid expansion of the gas tears the magma apart, and the volcano erupts violently, producing great volumes of ash. If enough gas escapes, the volcano can produce a sticky, slow-moving lava flow. Flows travel only a short distance from the vent before they solidify. The volcano tends to grow both vertically and laterally, resulting in a cone shape with steep slopes. Stratovolcanoes are not as voluminous as shield volcanoes.

There are dramatic differences in eruptions of Hawaiian volcanoes like Kilauea and Mount St. Helens. The different abundances of elements in magma, especially silica, exert the primary control on the explosiveness of an eruption. The viscosity of magma is greatly influenced by its silica content. Magmas which are low in silica tend to be very fluid. Most rocks in Hawaii are basalt. Basalts are characterized by a relatively low abundance of silica and high abundances of iron and magnesium. In contrast, most volcanic rocks along continental margins are andesite or dacite. Andesite or dacite are characterized by a relatively high abundance of silica and low abundances of iron and magnesium. Because Hawaiian magma is fluid, gas dissolved in the magma can escape prior to the eruption. In contrast, large amounts of gas is trapped inside andesitic or dacitic magmas. The gas cannot escape until the magma enters the throat of the volcano. When magma nears the vent, the gas bubbles nucleate and grow. The outward pressure exerted by the bubbles is greater than the strength of the magma. The lava fragments and is ejected violently at high velocity.

Conceptual image:



Decorative image:

Inference test:

Volcanoes test

Where is the least likely place for a volcano?

A. in the middle of a continent
B. at the edge of an ocean
C. on islands
D. under the ocean

What happens where plates diverge?

A. a trench forms that subducts oceanic crust
B. earthquakes
C. violent eruptions
D. new crust is formed

Which is true of converging oceanic and continental plates?

A. the oceanic plate is pushed deep into the mantle
B. they are generally free of earthquakes
C. continental plates are denser than oceanic plates
D. the two plates push up on each other and form mountains

What causes violent volcanic eruptions?

A. fluid magmas that are low in silica
B. magmas that come from melted continental plates
C. magmas that are high in basalt
D. magmas that come from melted oceanic plates

Which does not cause the creation of volcanoes?

A. oceanic–continental plate convergence
B. oceanic–oceanic plate convergence
C. continental–continental plate convergence
D. magma rising to the earth's surface

Memory test:

Volcanoes test

What magnitude earthquake accompanied the Mt. St. Helens eruption?

A. 2.3
B. 4.2
C. 5.1
D. 7.2

How many of the world's volcanoes are located on the perimeter of the Pacific Ocean?

A. none
B. about a third
C. over half
D. almost all

How many plates make up the earth's crust?

A. 2
B. 7
C. 12
D. about 20

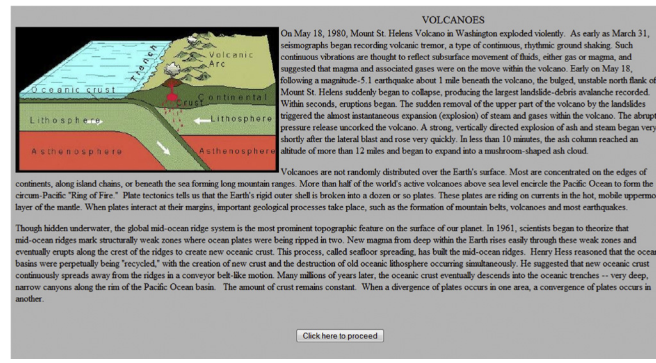
What is true of shield volcanoes?

A. they have steep sides
B. they are the largest
C. they erupt violently
D. they are also known as stratovolcanoes

What is true of andesitic magma?

A. it contains low amounts of silica
B. it contains low amounts of sulfur
C. it contains high amounts of magnesium
D. it contains high amounts of gas

Example screenshot



References

Ainsworth, S., & Burcham, S. (2007). The impact of text coherence on learning by self-explanation. *Learning and Instruction, 17*, 286–303. http://dx.doi.org/10.1016/j.learninstruc.2007.02.004.

Ainsworth, S., & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27*, 669–681. http://dx.doi.org/10.1207/s15516709cog2704_5.

Anderson, M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica, 128*, 110–118. http://dx.doi.org/10.1016/j.actpsy.2007.10.006.

Armbruster, B. B. (1984). The problem of "inconsiderate text". In G. G. Duffy, L. R. Roehler, & J. Mason (Eds.), *Comprehension instruction: Perspectives and suggestions* (pp. 202–217). New York: Longman.

Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In *The Cambridge handbook of multimedia learning* (pp. 135–146). http://dx.doi.org/10.1017/CBO9780511816819.009.

Baker, J. M., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review, 13*, 60–65. http://dx.doi.org/10.3758/BF03193813.

Butcher, K. R. (2006). Learning from text with diagrams: promoting mental model development and inference generation. *Journal of Educational Psychology, 98*, 182–197. http://dx.doi.org/10.1037/0022-0663.98.1.182.

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction, 20*, 155–166. http://dx.doi.org/10.1016/j.learninstruc.2009.02.014.

Chi, M. T. H. (2000). Self-explaining expository texts: the dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 5, pp. 161–238). Mahwah, NJ: Lawrence Erlbaum Associates.

Chi, M., Deleeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477. http://dx.doi.org/10.1016/0364-0213(94)90016-7.

Dunlosky, J., & Lipko, A. (2007). Metacomprehension: a brief history of how to improve its accuracy. *Current Directions in Psychological Science, 16*, 228–232. http://dx.doi.org/10.1111/j.1467-8721.2007.00509.x.

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic–extrinsic distinction of judgments of learning (JOLs): effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1180–1191. http://dx.doi.org/10.1037/0278-7393.27.5.1180.

Dunlosky, J., Rawson, K. A., & Middleton, E. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*, 551–565. http://dx.doi.org/10.1016/j.jml.2005.01.011.

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: why understanding students' judgments can improve their achievement. *Learning and Instruction, 24*, 58–61. http://dx.doi.org/10.1016/j.learninstruc.2012.05.002.

Garner, R., Brown, R., Sanders, S., & Menke, D. J. (1992). "Seductive details" and learning from text. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 239–254). Hillsdale, NJ: Erlbaum.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 702–718. http://dx.doi.org/10.1037/0278-7393.11.1-4.702.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*(1), 84–93. http://dx.doi.org/10.3758/BF03197714.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*, 1001–1013. http://dx.doi.org/10.3758/MC.37.7.1001.

Griffin, T. D., Wiley, J., & Salas, C. (2013). Supporting effective self-regulated learning: the critical role of monitoring. In R. Azevedo, & V. Aleven (Eds.), *International handbook of metacognition and learning technologies*. Springer Sciencehttp://dx.doi.org/10.1007/978-1-4419-5546-3_2.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*, 93–103. http://dx.doi.org/10.3758/MC.36.1.93.

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*, 414–434. http://dx.doi.org/10.1037/0022-0663.90.3.414.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*(6), 717–742. http://dx.doi.org/10.1006/jmla.1993.1036.

Hegarty, M., & Kozhevnikov, M. (1999). Types of visual–spatial representations and mathematical problem solving. *Journal of Educational Psychology, 91*, 684–689. http://dx.doi.org/10.1037/0022-0663.91.4.684.

Hegarty, M., & Sims, V. K. (1994). Individual differences in mental animation during mechanical reasoning. *Memory & Cognition, 22*(4), 411–430. http://dx.doi.org/10.3758/BF03200867.

Hegarty, M., Smallman, H. S., & Stull, A. T. (2012). Choosing and using geospatial displays: effects of design on performance and metacognition. *Journal of Experimental Psychology: Applied, 18*, 1–17. http://dx.doi.org/10.1037/a0026625.

Henderson, G. (1999). Learning with diagrams. *Australian Science Teachers Journal, 45*(2), 17–25.

Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research, 60*, 549–571. http://dx.doi.org/10.3102/00346543060004549.

Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading expository writing. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 239–254). Hillsdale, NJ: Erlbaum.

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

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: a theoretical analysis. *Journal of Memory and Language, 29*, 133–159. http://dx.doi.org/10.1016/0749-596X(90)90069-C.

Koriat, A. (1997). Monitoring one's own knowledge during study: a cue utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370. http://dx.doi.org/10.1037/0096-3445.126.4.349.

Lee, V. R. (2010). Adaptations and continuities in the use and design of visual representations in US middle school science textbooks. *International Journal of Science Education, 32*, 1099–1126. http://dx.doi.org/10.1080/09500690903253916.

Levin, J. R. (1981). On functions of pictures in prose. In F. J. Pirozzolo, & M. C. Wittrock (Eds.), *Neuropsychological and cognitive processes in reading* (pp. 203–228). New York: Academic Press. http://dx.doi.org/10.1016/B978-0-12-185030-2.50013-5.

Lewalter, D. (2003). Cognitive strategies for learning from static and dynamic visuals. *Learning and Instruction, 13*, 177–189. http://dx.doi.org/10.1016/S0959-4752(02)00019-1.

Linn, M. C., & Hsi, S. (2000). *Computers, teachers, peers: Science learning partners*. Mahwah, NJ: Erlbaum.

van Loon, M. H., de Bruin, A. B., van Gog, T., van Merriënboer, J. J., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154. http://dx.doi.org/10.1016/j.actpsy.2014.06.007.

Lowe, R. K. (1999). Extracting information from an animation during complex visual learning. *European Journal of Psychology of Education, 14*, 225–244. http://dx.doi.org/10.1007/BF03172967.

Lowe, R. K. (2003). Animation and learning: selective processing of information in dynamic graphics. *Learning and Instruction, 13*, 247–262. http://dx.doi.org/10.1016/S0959-4752(02)00018-X.

Lowe, R. K. (2004). Interrogation of a dynamic visualization during learning. *Learning and Instruction, 14*, 257–274. http://dx.doi.org/10.1016/j.learninstruc.2004.06.003.

Magliano, J. P., Todaro, S., Millis, K., Wiemer-Hastings, K., Kim, H. J., & McNamara, D. S. (2005). Changes in reading strategies as a function of reading training: a comparison of live and computerized training. *Journal of Educational Computing Research, 32*, 185–208. http://dx.doi.org/10.2190/1LN8-7BQE-8TN0-M91L.

Maki, R. H. (1998a). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–145). Hillsdale, NJ: Erlbaum.

Maki, R. H. (1998b). Predicting performance on text: delayed versus immediate predictions and tests. *Memory & Cognition, 26*, 959–964. http://dx.doi.org/10.3758/BF03201176.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 663–679. http://dx.doi.org/10.1037/0278-7393.10.4.663.

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(4), 609–616. http://dx.doi.org/10.1037/0278-7393.16.4.609.

Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review, 1*, 126–129. http://dx.doi.org/10.3758/BF03200769.

Maki, R. H., & Serra, M. (1992). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology, 84*(2), 200–210. http://dx.doi.org/10.1037/0022-0663.84.2.200.

Mayer, R. E. (1993). Illustrations that instruct. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 4, pp. 253–284). Hillsdale, NJ: Lawrence Erlbaum.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Presshttp://dx.doi.org/10.1017/CBO9781139164603.

Mayer, R. E. (Ed.). (2005). *Cambridge handbook of multimedia learning*. New York, NY: Cambridge University Presshttp://dx.doi.org/10.1017/CBO9780511816819.

Mayer, R. E., Bove, W., Bryman, A., Mars, R., & Tapangco, L. (1996). When less is more: meaningful learning from visual and verbal summaries of science textbook lessons. *Journal of Educational Psychology, 88*, 64–73. http://dx.doi.org/10.1037/0022-0663.88.1.64.

McNamara, D. S. (2004). SERT: self-explanation reading training. *Discourse Processes, 38*, 1–30. http://dx.doi.org/10.1207/s15326950dp3801_1.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*, 349–363. http://dx.doi.org/10.1037/0096-3445.131.3.349.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*(3), 159–163. http://dx.doi.org/10.1111/j.1467-8721.2009.01628.x.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*(4), 530–542. http://dx.doi.org/10.1037/0096-3445.132.4.530.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133. http://dx.doi.org/10.1037/0033-2909.95.1.109.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: the "delayed-JOL effect". *Psychological Science, 2*(4), 267–270. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00147.x.

Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology: Section A, 55*(2), 505–524. http://dx.doi.org/10.1080/02724980143000352.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*, 1004–1010. http://dx.doi.org/10.3758/BF03209348.

Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning & Instruction, 22*, 262–270. http://dx.doi.org/10.1016/j.learninstruc.2011.10.007.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological Bulletin, 137*(1), 131–148. http://dx.doi.org/10.1037/a0021705.

Salomon, G. (1984). Television is "easy" and print is "tough": the differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology, 76*(4), 647–658. http://dx.doi.org/10.1037/0022-0663.76.4.647.

Sanchez, C. A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition, 34*, 344–355. http://dx.doi.org/10.3758/BF03193412.

Sanchez, C. A., & Wiley, J. (2009). To scroll or not to scroll: scrolling, working memory capacity, and comprehending complex texts. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 51*, 730–738. http://dx.doi.org/10.1177/0018720809352788.

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 49–69). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816819.005.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgments reflect the belief that diagrams improve learning from text. *Memory, 18*, 698–711. http://dx.doi.org/10.1080/09658211.2010.506441.

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*, 129–160. http://dx.doi.org/10.1016/S0361-476X(02)00011-5.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73. http://dx.doi.org/10.1037/0022-0663.95.1.66.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1267–1280. http://dx.doi.org/10.1037/0278-7393.31.6.1267.

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*, 331–362. http://dx.doi.org/10.1080/01638530902959927.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition and self-regulated learning* (pp. 85–106). New York: Routledge.

Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2012). Elementary school experience with comprehension testing may influence metacomprehension accuracy among 7th and 8th graders. *Journal of Educational Psychology, 104*, 554–564. http://dx.doi.org/10.1037/a0028660.

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*, 264–273. http://dx.doi.org/10.1348/135910710X510494.

Tversky, B. (1995). Cognitive origins of conventions. In F. T.Marchese (Ed.), *Understanding images* (pp. 29–53). New York: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4613-8380-2_4.

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 79–111). Cambridge: MIT Press.

Tversky, B., Zacks, J., Lee, P., & Heiser, J. (2000). Lines, blobs, crosses and arrows: diagrammatic communication with schematic figures. In *Theory and application of diagrams* (pp. 221–230). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/3-540-44590-0_21.

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology, 132*, 408–428. http://dx.doi.org/10.3200/GENP.132.4.408-428.