

Metacognitive Monitoring During and After Reading

Keith W. Thiede

Boise State University

Thomas D. Griffin & Jennifer Wiley

University of Illinois at Chicago

Joshua S. Redford

Boise State University

running head: Metacomprehension

Send correspondence to: Keith Thiede

Boise State University

Education

1910 University Drive

Boise, ID 83725-1745

KeithThiede@boisestate.edu

Abstract

Models of self-regulated learning suggest that accurate monitoring is critical to effective regulation of study, and judgment accuracy has been empirically linked to successful learning. Therefore, to improve students' ability to learn from text, it is important to discover ways to improve monitoring accuracy during and after reading. Although accuracy for monitoring performance can be quite high in associative learning tasks, the accuracy for monitoring comprehension during or after reading-to-learn tasks, which has been called metacomprehension accuracy, has typically been quite low. In this chapter, we will review research on metacognitive monitoring and describe factors that may constrain metacomprehension accuracy. We will then review recent research that has addressed these constraints and produced unprecedented levels of monitoring accuracy during or after reading. Finally, future research directions will be discussed.

Metacognitive Monitoring During and After Reading

Models of self-regulated learning (e.g., Butler & Winne, 1995; Metcalfe, 2002; Nelson & Narens, 1990; Thiede & Dunlosky, 1999) describe learning as a process that involves setting a learning goal, monitoring progress toward that goal, and regulating study in order to achieve that goal. For example, consider a student preparing for an upcoming test. As the student studies, she monitors her progress toward the goal of mastering the material. If her monitoring indicates that she has not yet mastered the material, she will likely restudy the material until her monitoring suggests that the material has been mastered, at which time she will stop studying. Accurate metacognitive monitoring is critical to effective regulation of study (Thiede, 1999; Winne & Perry, 2000). If a person does not accurately differentiate well-learned material from less-learned material, she could waste time studying material that is already well learned or fail to restudy material that has not yet been adequately learned.

Thiede, Anderson, and Therriault (2003) demonstrated the important role that accurate monitoring plays in learning from text by manipulating monitoring accuracy and examining the effect on regulation of study and learning outcomes. In this experiment, an initial important finding was that participants who were given a delayed-keyword-generation instruction (i.e., asked to generate a list of five keywords that captured the essence of a text prior to judging comprehension) more accurately monitored their comprehension on a first set of tests than did participants in a no-keyword-generation or immediate-keyword-generation group. Next, all participants were given the opportunity to select texts for re-study. The participants in the delayed-keyword group made better decisions about which texts to reread than did participants in the other two groups. The mean proportion correct on first tests for the texts selected for rereading versus those not selected for rereading was .27 versus .78 respectively. The large

difference in performance between these sets of texts indicates that this group was able to accurately discriminate the texts they comprehended well, from those they did not, and specifically chose those texts with poor comprehension for restudy. By contrast, for the other two groups, the mean proportion correct on first tests for the texts selected for rereading versus those not selected for rereading was .43 versus .53 respectively. These participants were either less able to discriminate what they had learned well than the delayed-keyword group, or were less likely to use their discrimination as a basis for restudy selections. More important, the more effective regulation of study observed in the delayed-keyword group led to higher overall reading comprehension performance as assessed by a second set of tests. Thus, this study demonstrated that improving metacomprehension accuracy improved self-regulated study, and ultimately improved comprehension.

Improving Monitoring Accuracy

The majority of investigations exploring factors that affect the accuracy of metacognitive monitoring have been done using associative learning tasks (e.g., learning word pairs, translations or definitions). This research has shown that relative monitoring accuracy, as operationalized by computing intra-individual correlations (usually gamma correlations) between judgments of learning and test performance across items, improves when a person monitors learning (a) after a delay rather than immediately after studying an item (Dunlosky & Nelson, 1992, 1997; Kelemen & Weaver, 1997; Nelson & Dunlosky, 1991; Weaver & Kelemen, 1997), (b) after practice monitoring his or her learning (Vesonder & Voss, 1985), and (c) following a practice test of the material (King, Zechmeister, & Shaughnessy, 1980; Lovelace, 1984;

Shaughnessy & Zechmeister, 1992). Furthermore, under certain circumstances, people can attain near perfect levels of monitoring accuracy. For example, monitoring accuracy was $+0.93$ for delayed judgments of learning in Nelson and Dunlosky (1991).

In sharp contrast to accuracy of monitoring during associative learning tasks, metacomprehension accuracy has typically been quite low. Standard metacomprehension paradigms use measures of predictive, relative accuracy from expository texts. Similar to metamemory approaches, readers are asked to read a set of texts, make judgments of learning or predictions about their performance on future tests, and then complete the tests. Again, intra-individual correlations (traditionally gamma correlations) are computed from this, and the average hovers around $.27$. Maki (1998a) reported this figure as the mean intra-individual correlation between comprehension ratings and test performance across 25 studies from her lab. A recent analysis by Dunlosky across his lab's studies arrived at the same exact figure (Dunlosky & Lipko, 2007). Other reviews by Lin and Zabrocky (1998) and Weaver, Bryant, and Burns (1995) reached similar conclusions

Factors Contributing to Poor Monitoring Accuracy when Learning from Text

Monitoring learning from text may be a task constrained by rather different factors than monitoring learning of paired-associates (Wiley, Griffin & Thiede, 2005). Consider the task of judging learning in an associative learning task. A person studies a list of word pairs (e.g., dog – spoon) and is instructed that the test will involve recalling the second word when given the first. After studying the list of word pairs, the person is given the first word and asked to judge how confident he is that he will recall the second word in the future. In this case, the person should

be perfectly clear about what he is being asked to judge, and this expectation maps perfectly onto the test. As noted above, delayed judgments of learning (JOLs) can produce nearly perfect monitoring accuracy, particularly when the prompt for judgments is the stimulus-alone cue (cf. Dunlosky & Nelson, 1992, who showed a large drop in accuracy when the judgment is prompted with a stimulus-response pair).

Contrast this with the process of judging learning from text. In the original Glenberg and Epstein (1985) paradigm, a person read a series of sixteen 200 – 400 word texts. The person then rated his comprehension for each text on a 7-point scale, and answered one inference question per text to provide a measure of comprehension. Metacomprehension accuracy was computed for each individual by correlating judgments of comprehension and test performance across the texts.

As noted above, such accuracy is generally observed to be quite poor, and several potential reasons can be gleaned by noting differences between learning from text and paired associates paradigms. These differences largely stem from the inherently greater complexity and vagueness of what it means to comprehend a text versus what it means to recall a word when cued by its associate. In the remainder of this chapter, we will review research that focuses on several obstacles to effective monitoring while learning from text and potential solutions that have been offered to address the complexity and vagueness of the monitoring task. First, we briefly describe these constraints, then we explore them more fully in light of interventions designed to overcome them.

Some of the blame for poorer accuracy and null or unreliable findings in metacomprehension compared to metamemory research probably lies with the lower validity in the performance measures of text comprehension versus cued word recall. Actually, research on

“monitoring accuracy” does not solely measure monitoring accuracy, but rather measures judgment-performance covariance, which is just as influenced by the validity of the performance measures as by the accuracy of learners’ judgments. The validity of our inferences about monitoring accuracy depends on the validity of the performance measures that greatly constrain any judgment-performance covariance. Metamemory research largely avoids this problem, because a valid recall measure requires little more than assessing recall for each word–pair that was judged. However, texts and the comprehension of them are so complex and multi-dimensional that it is difficult to adequately define the construct, let alone measure it for a particular text. One implication is that we must carefully consider variability in the validity of our performance measures before drawing too many inferences about changes or differences in monitoring accuracy between conditions, studies, or paradigms. We should also take steps to improve the validity of our comprehension measures.

With comprehension measures, a major validity issue is the one of content coverage. Weaver (1990) discussed the problem that a one-item test does not provide a valid or reliable measure of comprehension, and demonstrated that metacomprehension accuracy improves when multiple comprehension items are given at test. The failure to use comprehension tests with multiple items per text is one likely reason for the low correlations observed in several early studies in the metacomprehension literature (i.e. Glenberg & Epstein, 1985; Glenberg, Sanocki, Epstein & Morris, 1987, Experiment 1).

Beyond issues of reliability, Wiley et al. (2005) expanded on the need for multiple comprehension items per text by arguing that any test that lacks complete coverage of the text information will lack validity and tend to under-estimate true monitoring accuracy levels. What counts as coverage depends on the units of text being judged. Whether readers are asked to judge

a whole text or certain portions, there are components and their inter-relations that contribute to the unit of the text being judged. Comprehension of one component may correlate only modestly with other components. Thus, coverage of all the various components and their relations within a unit of text is a minimal requirement for the validity of a measure intended to assess comprehension of a text unit.

Without complete coverage, judgment-performance alignment will not only be unreliable but will systematically under-estimate true monitoring accuracy. For instance, imagine a person who accurately knows that he understood 50% of a text, so he gives a mid-range judgment. If the test items only constitute 50% coverage, then his performance-judgment alignment will be misleadingly poor, regardless of whether the test taps the 50% he knew (actual performance being higher than judged), or the 50% he did not know (actual performance being lower than judged). The only situation where incomplete coverage would overestimate a reader's accuracy is the low-probability scenario where the reader bases their judgment on only a portion of the text and that happens to be the same portion tapped by the test items. Performance measure validity is a problem to be considered and overcome, but there are other factors that constrain the actual rather than just computed accuracy of learners' judgments.

Dunlosky and Lipko (2007) address a different form of coverage problem. They argue that poor metacomprehension accuracy is a function of judgments that cover too much text material. They point out that grain size of the typical monitoring judgment (which is global and covering the whole text) is much larger than the grain size of the test items (which may concern specific concepts from within the text). Readers may struggle to compute a judgment covering many components of the text. This is not a problem in typical paired-associate tasks, because the judgment and performance grain sizes are identical and at the level of associated pairs.

Yet another factor that may contribute to poor monitoring accuracy when learning from text is a poor alignment between which cues a person uses for metacomprehension judgments and which cues predict performance on a test of comprehension. The notion that assessments can vary in the level of representation they tap is derived from Kintsch's theory of comprehension, which posits that text is mentally represented in multiple forms from surface memory of the words to a logically integrated conceptual model (Kintsch, 1994). For instance, a mismatch can occur if a person judges comprehension based on their ability to recall the words from the text, but the test assesses whether they can apply the concepts to a new scenario (Rawson, Dunlosky, & Thiede, 2000; Wiley et al., 2005). Again, paired-associate metamemory tasks sidestep this problem. For the highly accurate delayed-JOLs, the judgment prompt is identical to the test prompt; thus, there is no possible misalignment. Even when the judge and test prompts differ, paired-associate learners still need to only monitor their memory for the words and not a potentially orthogonal level of conceptual understanding.

A final factor may be the differential demands of the metacomprehension versus the metamemory monitoring process. Metacomprehension monitoring more so than metamemory monitoring may require concurrent attention at two levels of processing: directing attention to the monitoring of cues plus directing attention to processing the text. The inherent complexity of text processing and the comprehension process seem to entail that most valid cues will only be available at the time of judgment if they were attended to during reading itself. Based on this assumption, poor metacomprehension accuracy may occur especially among low ability readers due to high concurrent processing requirements (Griffin, Wiley & Thiede, 2008). Yet again, delayed-JOL metamemory tasks do not pose this obstacle when the judgment prompt is identical

to the test prompt, because the judgment prompt itself gives rise to many if not most cues that learners need to predict future test performance.

To provide a current overview of work in the metacomprehension literature, Table 1 presents all studies that have been done in the area. The table is restricted to published experimental papers that use a predictive, relative accuracy paradigm with expository texts. For each study, the mean intra-individual correlation (γ) for each condition is reported. As noted above, due to issues of reliability of tests with only 1 or 2 items (Weaver, 1990) only studies with more than two items per test are included. Interestingly, an average of the “standard” predictive accuracy conditions in this table also leads to an average γ of .27. However, there are also several recent studies which have produced substantially higher levels of accuracy, and these studies form their own distribution toward the top end of the range. The second half of this chapter will describe how several lines of recent research have provided evidence for the various constraints discussed above, and how they may be overcome to improve monitoring accuracy for texts to substantially higher levels.

Aligning the Grain-size of Judgments and Tests: Term-specific Judgments

Metacomprehension judgments usually ask individuals to make a global assessment of their text comprehension—a general evaluation of their overall understanding of the material. Although individuals may be sensitive to their level of comprehension for specific concepts, it is unclear whether they would be able to accurately translate this into a single value. Moreover, criterion tests often involve testing people's learning of specific concepts within a passage, which also may not map cleanly onto a global judgment. Thus, the use of a single global judgment and

the mismatch in grain size between monitoring judgments and subsequent tests provides one potential explanation for the low monitoring accuracy reported in previous research (Dunlosky & Lipko, 2007).

Dunlosky, Rawson, and Middleton (2005) tested the possibility that a mismatch in grain size undermines accurate measurements of monitoring accuracy while learning from text by reducing the grain size of the monitoring judgments. In addition to asking participants to make a global judgment of their level of comprehension, participants also made monitoring judgments regarding their ability to recall the meaning of specific terms contained in the text. Surprisingly, the accuracy of these term-specific memory judgments was not dramatically different than the accuracy of the global comprehension judgment.

Dunlosky et al. (2005) discussed two issues that might prevent a concordance in grain size from improving monitoring accuracy when learning from text. First, participants were relying on their familiarity with specific terms rather than attempting to recall term definitions prior to making their judgments. When forced to attempt recall of definitions prior to making judgments, accuracy increased to .73 (see also Dunlosky, Rawson & McDonald, 2002, for a similar result). Second, in agreement with Koriat's accessibility model (1993), participants who attempted recall prior to their judgments partly based their judgments on the overall quantity of retrieved information rather than on the quality of the retrieved information. In a subsequent study, Rawson and Dunlosky (2007) found that participants who received feedback regarding the accuracy of their recalled definitions (the original definition from the text being displayed next to their typed definition) prior to their judgments further improved their monitoring accuracy to .92.

Note that the grain-size approach also addresses the content coverage issue discussed above. Thus, another way that grain size may be helping the concordance between predicted

performance and actual performance on the test is by making the coverage of the test obvious. As we discuss below, the nature of the test (in this case, a test of memory for specific definitions) also is made explicitly clear to learners through this procedure. A final way that this procedure may aid accuracy is through providing the exact test items, and giving students the correct answers to the exact test items. With this manipulation, the judgment that is made changes from a prediction to a postdiction, and thus it becomes more similar to the conditions that promote high levels of performance in paired-associate paradigms, and replicates the robust effects that have been found across many studies that postdictions are superior to predictions of metacomprehension accuracy (Glenberg & Epstein, 1985; 1987; Glenberg, Sanocki, Epstein & Morris, 1987; Lin, Moore & Zabrocky, 2001; Maki, Foley, Kajer, Thompson & Willert, 1990; Maki & Serra, 1992a, 1992b; Pierce & Smith, 2001). Nevertheless, this research demonstrates that a mismatch in grain size can produce lower levels of monitoring accuracy when learning from text. When grain size is matched and other risk factors are accounted for, monitoring accuracy can increase dramatically, and the present levels are well beyond the magnitude of the improvements that are typically seen as a result of exposure to the test items.

Aligning the Level of Understanding Assessed by Judgments and Tests

Another factor undermining monitoring accuracy when learning from text pertains to the perspective individuals adopt when reading a text and assessing their level of text comprehension (Wiley et al., 2005). Most metacomprehension research has employed expository texts and asked students to judge their level of comprehension without further specification of what the construct of comprehension might mean. Although meaningful comprehension of expository texts requires

an appreciation of the information at a deeper level (e.g., how or why some event occurs), many students approach learning from texts in the same way as they approach learning word lists—as a collection of discrete pieces of information to be stored for later retrieval. To the extent that comprehension tests are more concerned with a reader’s ability to make inferences, connections or develop an understanding of how or why some event or phenomenon occurs, then metacomprehension judgments may be inaccurate because individuals are basing their judgments on cues (e.g., memory of details) that are not predictive of test performance (i.e., the ability to make inferences). Thus, a second explanation for the low monitoring accuracy reported in previous research is the mismatch between the kinds of cues that readers use to predict their own comprehension, and the kinds of cues that they should use.

When assessing comprehension, a person can monitor various cues that are produced by comprehension processes, such as the fluency of text processing (Dunlosky & Rawson, 2005; Rawson & Dunlosky, 2002; Lefevre & Lories, 2004; Maki, Shields, Wheeler & Zacchilli, 2005; Rawson, Dunlosky & McDonald, 2002) or their familiarity with the topic (Maki, 1998a; Jee, Wiley & Griffin, 2006) or ability to remember parts of the text (Thiede, Griffin, Wiley, & Anderson, in press). A judgment of comprehension is then based on an inference about those cues (Schwartz, Benjamin, & Bjork, 1998). In the case of fluency, people presumably judge that their understanding of a text is better when the text is easily processed than when it is difficult to process (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989). Similarly, people may judge that their understanding of a text is better when elements of the text are easily recalled, or when they are interested in or familiar with the topic. According to this cue-utilization framework of metacognitive monitoring (Koriat, 1997), the accuracy of comprehension judgments will be a function of the degree to which the cues used in assessing comprehension empirically correlate

with performance on comprehension tests. More specifically, assuming the inferences are valid, metacomprehension accuracy will increase as the available cues more highly correlate with subsequent test performance. The particular problem in the case of comprehension judgments is that readers often do use fluency, topic and memory-based cues to judge their understanding, but these cues are not necessarily predictive of actual comprehension (Thiede, Griffin, Wiley, & Anderson, in press). To determine the types of cues that are highly predictive of performance on tests of comprehension, it is necessary to understand the comprehension processes that provide many of those cues. According to the construction-integration model of comprehension (Kintsch, 1998), readers construct meaning from text at several levels of representation: a lexical or surface level, a textbase level, and a situation model level. The lexical level, containing the surface features of the text, is constructed as the words and phrases appearing in the text are encoded. The textbase level is constructed as segments of the surface text are parsed into propositions, and as links between text propositions are formed based on argument overlap and other text-explicit factors. However, the deeper understanding of the text, and the scenario it describes, is constructed at the level of the situation model, which involves connecting text information with the reader's prior knowledge and using it to generate inferences and implications from the text. It is a reader's situation model that largely determines his or her performance on tests of comprehension (McNamara, Kintsch, Songer, & Kintsch, 1996). Therefore, getting people to base their judgments on cues related to their situation model rather than their surface model or textbase should increase the predictive accuracy of judgments when tests are comprehension-based (Rawson et al., 2000; Wiley et al., 2005).

This situation model approach has been supported by a number of studies that have improved metacomprehension accuracy through manipulations designed to prompt readers to

access and utilize their situation models during or after reading. These manipulations have included generating summaries or keywords after a delay (Anderson & Thiede, 2008; Thiede & Anderson, 2003; Thiede et al., 2003; Thiede, Dunlosky, Griffin & Wiley, 2005), engaging in self-explanation while reading (Griffin et al., 2008; Wiley, Griffin & Thiede, 2008), and engaging in concept mapping while reading (Thiede et al., in press).

Delayed-generation effect. In several studies, Thiede and his colleagues have found that having students engage in delayed generation tasks produced unprecedented levels of metacomprehension accuracy (Anderson & Thiede, 2008; Thiede & Anderson, 2003; Thiede et al., 2003; Thiede et al., 2005). In Thiede and Anderson (2003), metacomprehension accuracy was dramatically greater for a delayed-summary group (mean correlation around .60) than for other groups (mean correlation around .26) across two experiments (one with shorter texts and one with longer texts).

Thiede et al. (2003) showed that generating a list of five keywords that captured the essence of a text produced a similar boost in metacomprehension accuracy (mean correlation of .71) as had generating summaries. Moreover, the effect was again influenced by the timing of generation. That is, generating keywords after a delay produced dramatic improvement, whereas generating keywords immediately after reading did not affect accuracy compared to a control group (mean correlation of both around .25).

As noted above, the cue-utilization framework of metacognitive monitoring (Koriat, 1997) and the construction-integration model of comprehension (Kintsch, 1998) provide a framework to explain why delayed generation tasks improve metacomprehension accuracy.

Writing a summary or generating keywords may allow a reader to reflect on how successfully he or she can retrieve information during the generation task (cf. the modified

feedback hypothesis described by Glenberg et al., 1987). Accordingly, a text may receive a high rating of comprehension if the person was able to retrieve a great deal of information about the text during the generation task; whereas, a text may receive a low rating of comprehension if the person struggled to retrieve information about the text during summarization or keyword generation.

Further, the timing of the generation task is critical. Kintsch, Welsch, Schamlhofer, and Zimny (1990) showed that memory for surface and textbase representations of text decays over time, whereas the situation model is robust to such decay. Thus, when writing a summary or generating keywords immediately after reading, a person may have easy access to the lexical and textbase representations and the person can use this information for generation. However, feedback gained during immediate generation tasks may not provide an accurate basis for judgments of comprehension because performance on the generation tasks is being determined largely by the surface and textbase representations. In contrast, when writing a summary after a delay, the findings by Kintsch et al. suggest that a person will likely have access primarily to the situation model for the text. Thus, using delayed summary and keyword generation tasks provides more valid cues for judgment of comprehension because both the comprehension judgments and test performance are based on the situation model.

A stronger test of the situation model approach was to determine which delays might be responsible for the delayed generation task effect observed in previous studies (Thiede & Anderson, 2003; Thiede et al., 2003). According to the situation model approach, the critical delay for better metacomprehension should be the one between reading and summarizing or keyword generation because this delay causes readers to access their situation model prior to judgment. The generation task forces readers to access and use their text representation, and

doing so at a delay after reading means that their representation is more likely be based on a situation model than the surface or textbase level since they should have decayed during the delay. Although this interpretation fits the findings of Thiede and Anderson, and Thiede et al., 2003, there were actually several features of the previous studies' designs and procedures that were confounded and preclude them from providing clear support for the situation model approach. In Thiede et al (2003), the delayed condition was created by having readers generate keywords for all texts after reading all the texts, rather than immediately after each text. This produced not only a delay between reading and keyword generation, but also decreased the time lag between keyword generation and judgments of comprehension, and decreased the time lag between each keyword generation by having all generation tasks performed in succession for the delayed group. In contrast to the reading-keyword delay, these other two time lags could account for the superior accuracy in the delay group without implying anything about access to the situation model.

To test for the viability of this alternative explanation, Thiede et al. (2005) manipulated the possible effects of different lags on metacomprehension accuracy. As altering the order of tasks in the protocol will always affect more than one time lag, a set of two experiments was constructed, where each one evaluated the possible effects of one of the confounding lags. Both experiments replicated the earlier findings and showed that the delay between reading and generating keywords was critical for improving metacomprehension accuracy. Experiment 1 also showed that the keyword-judgment lag was not critical, while Experiment 2 showed that the lag between multiple keyword tasks was not critical. In addition, Experiments 3 and 4 showed that the critical factor was actually generating keywords at a delay and not merely thinking about the text or being re-exposed to key ideas of the texts at a delay.

Thus, this set of four experiments provided a compelling case that the act of accessing and making use of one's representation via a generation task at a delay after reading are the critical factors that have produced some of the largest improvements in metacomprehension accuracy observed to date. Moreover, these findings provide strong support for the situation model approach. They suggest that it is access to valid cues (i.e., those based in a situation model) that underlies why both the reading-keyword delay and the generation component of the keyword task combine to produce the observed improvements in metacomprehension accuracy.

Self-explanation and cue-access. The previous findings suggest that getting readers to focus on their situation model during monitoring will improve metacomprehension accuracy. Another way to test this hypothesis is to examine interventions that have been used to get readers to construct more complete and coherent situation models. Such interventions should more generally increase readers' attention to their situation-model representations, giving them greater access to cues about the quality of their situation-model that they could use to more accurately predict their comprehension performance.

One such intervention that has improved situation-model construction is self-explanation. Instructing and/or training readers to explain the logical and causal connections among idea units while they read has been shown to improve the completeness and coherence of the mental models readers construct of complex causal phenomena (e.g., Chi, DeLeeuw, Chiu, & Lavancher, 1994; Hemmerich & Wiley, 2002; McNamara, 2004; Wiley & Voss, 1999; Wiley, 2001). The interventions have ranged from simple global task instructions (Wiley & Voss, 1999; Wiley, 2001) to targeted prompts to explain specific relations (e.g., Chi et al., 1994), to more extensive recursive training and feedback designed improve the quality of self-explanations (e.g., McNamara, 2004). Based on these previous findings, a self-explanation instruction should

prompt readers to focus on and use their situation models during reading. If this is the case, the situation model approach would also predict that self-explanation tasks should improve the accuracy of metacomprehension judgments during reading.

A recent study done by Griffin et al. (2008) has provided support for this hypothesis. In this study, giving readers a one page pre-reading instruction that prompted them to self-explain as they read the expository texts increased metacomprehension accuracy to levels on par with those found for delayed-generation tasks (mean correlation around .60) as compared to .39 in control conditions. Wiley, Griffin and Thiede (2008) recently found a similar result with a similar self-explanation condition. Another finding by Thomas and McDaniel (2007) may also be related, as they found improved monitoring accuracy specifically on a conceptual (rather than detail-oriented) test following a reading task that required the re-ordering sentences (rather than inserting missing letters). This sentence-sorting manipulation could also have improved attention to the situation-model level.

A key difference between these studies and delayed-generation tasks used in previous studies is that the boost in accuracy occurred even without introducing a delay between reading and judging comprehension. Besides addressing potential pragmatic pedagogical concerns, this avoids an alternative explanation for the delayed-generation effects, namely a transfer-appropriate-monitoring account which would posit that increased accuracy is due to the fact that the processing during the generation task and the target test both occur at a delay. Instead, the self-explanation results provide specific support for the situation model approach, that metacomprehension accuracy will improve when readers are made to focus on valid cues for predicting comprehension, rather than using cues based on familiarity, fluency or surface memory for text.

Concept mapping and cue access. An intervention similar to self-explanation that has been used in the literature, especially with younger or less-skilled readers, is concept mapping. Weinstein and Mayer (1986) suggested that instructing students to create concept maps of texts during reading helps them to identify the connections among concepts in a text. In discussion of concept mapping as an intervention, Weinstein and Mayer suggest that argumentation, self-explanation, and concept mapping tasks all help readers to construct and pay attention to the underlying causal, situation models of the subject matter. As the text is available during the activity, concept mapping tasks may be especially appropriate for less-able readers who may have difficulty engaging in a task that requires them to remember a text (Nesbit & Adesope, 2006; Stensvold & Wilson, 1990).

Using a within-subjects design, Thiede, Griffin, Wiley, and Anderson (in press) gathered initial metacomprehension data by running participants through the standard experimental procedure for assessing metacomprehension accuracy. That is, participants read a series of texts, judged their comprehension of each text, and then completed a comprehension test for each text. Participants, who were college students in a remedial reading course, then received eight days of instruction on how to construct concept maps for texts. They then completed the standard procedure for assessing monitoring accuracy again, but this time they constructed concept maps while reading the new set of texts. As with the self-explanation instructions, constructing concept maps while reading increased students' metacomprehension accuracy to around .65, which was a significant improvement from pre-intervention levels of around .32. These findings provide further support for the basic tenet of the situation model approach that getting readers to access and use their situation model prior to judging comprehension improves their

metacomprehension accuracy. And again, this boost in accuracy occurred without introducing a delay between reading and judgments.

The interventions discussed up to now all have dealt with increasing the alignment (of either grain size or level of representation) between judgments and performance measures as readers attempt to learn from text. We now turn to an accuracy constraint that deals with the demands of the monitoring task itself.

Demands of Monitoring during Reading

Unlike monitoring metamemory of paired associates, monitoring metacomprehension of text requires concurrent attention to two cognitive levels: processing the text and monitoring the products of comprehension. These levels have been defined previously as the *object-level* where the incoming information such as a text is processed and a mental representation of it is formed, and the *meta-level* where cues about one's mental representation are the information being processed and monitored (see Fischer & Mandl, 1984; Nelson & Narens, 1990). The simplicity of cued recall in associative learning paradigms means that the judgment prompt is largely identical to the testing prompt. Learners can simply perform the actual memory test at the time of judgment and use any performance feedback to make fairly accurate judgments about future test performance. They can attend to the meta-level after much of the object-level processing, the reading and studying of the word-pair, is complete.

In contrast, text processing is more complex and involves many aspects of comprehension that typically are not (and cannot realistically be) fully contained in a judgment prompt. This is especially true of global judgments such as "How well did you understand that

text?” Thus, valid monitoring cues for text comprehension judgements cannot be produced at the time of judgment, and must be generated during the reading of the text. Access to those cues during later judgment will depend on how much attention was paid to them during the processing of the text. Thus, accurate monitoring of text comprehension requires that readers concurrently attend to both processing the text at the object-level and processing cues about their level of comprehension at the meta-level during reading. (Although with extremely short or conceptually simple texts, most cues produced during reading may remain highly available even after reading.)

Griffin et al. (2008) tested the hypothesis that having to concurrently attend to (and/or repeatedly switch between) two processing levels limits the accuracy of metacomprehension judgments. In this study, judgment accuracy was shown to be predicted by individual differences in two abilities (e.g., comprehension ability and working memory capacity) that are theoretically related to the ability to monitor meta-level cues while concurrently processing the text during reading. Moreover, a rereading manipulation demonstrated that these demands were alleviated and accuracy improved when readers were allowed to read the text twice prior to judgment. Benefits for rereading were found specifically for readers with poor comprehension ability, as well as for readers with low working memory capacity. This latter finding is especially telling as it implicates that poor monitoring is due to a more general inability to allocate or control attentional resources, which is the construct purportedly measured by working-memory span tasks (Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle 2005). Presumably, readers lower in either working memory or comprehension ability struggled with concurrently attending to both the meta-level and object-level during a first reading, but could devote less attention

towards text processing thus more towards meta-level cues and monitoring during the second reading.

This work differs from the rest reported in this chapter in that it deals with obstacles to accurate monitoring due to the demands of the monitoring process itself, which may exceed a reader's available resources, rather than obstacles due to mismatches in the nature of the cues that are used for judgment. It also extends other research that has demonstrated improvements in metacomprehension accuracy with rereading manipulations (Rawson et al., 2000) and specifically for immediate rather than delayed rereading (Dunlosky & Rawson, 2005). Although Rawson and colleagues argued that rereading can make situation-model based cues more available, the fact that Griffin et al.'s (2008) rereading effects were moderated by individual differences in attentional control, coupled with the fact that rereading effects were modest compared to self-explanation effects (a manipulation that we presume does direct attention to the situation model) suggests that rereading effects may depend more on freeing attentional resources than increasing the availability of situation-model based cues.

A final set of findings also seems to fit in with this proposed demands of monitoring constraint. The lone observation of predictive metacomprehension accuracy above .5 before the turn of the century was an interesting finding by Weaver and Bryant (1995) that high levels of metacomprehension accuracy can be observed when texts are moderately difficult (correlations between .66 and .69 across two experiments). Moderately difficult texts led to superior metacomprehension accuracy over either texts that were too easy, or texts that were too difficult in terms of their Flesch readability scores. A similar pattern using the same texts was replicated by Lin, Zabrocky and Moore (2002). Especially the finding that moderately difficult texts lead to better metacomprehension than difficult texts is consistent with the demands of monitoring

view because texts that are too difficult would leave no resources for comprehension monitoring. Explanations for the poor accuracy on easy texts have attributed poor performance to either a lack of discriminability among the easy set, or a lack of engagement (Dunlosky & Rawson, 2005; Weaver & Bryant, 1995)

Summary

Although traditionally levels of metacomprehension accuracy have been observed to be quite low, several studies in the last decade have demonstrated substantial increases in monitoring accuracy when readers are learning from text. The common thread that unites all of these attempts to improve monitoring during or after reading is that they address constraints that arise from the inherent complexity of monitoring learning from text. There are additional constraints and obstacles when learning involves text comprehension that are not faced when learning involves cued recall of paired associates. As such, this work may begin to provide insights into a puzzle that has faced metacognitive researchers for decades, namely why metacognitive accuracy suffers so greatly when the to-be-learned material goes from paired associates that must be recalled to texts that must be comprehended.

Most of the successful interventions presented here were based in a situation model approach to improving metacomprehension accuracy. It is important to note that the situation model approach does not predict that readers who form better situation models will necessarily have better metacomprehension (Griffin, et al., 2008; Jee, Wiley & Griffin, 2006). Improving the quality of the representation that can be constructed is not the point of these interventions. Instead what is important is that readers access and use the situation model to inform their

judgments of comprehension. Although self-explaining and concept-mapping may also improve situation-model construction and may ultimately lead to better comprehension, the goal of these interventions was to direct readers to attend to valid cues for comprehension.

It is also important to distinguish the present situation model approach, from an earlier approach in much the same spirit: the levels-of-disruption hypothesis (Dunlosky & Rawson, 2005; Dunlosky, Rawson & Hacker, 2002). The levels-of-disruption hypothesis shares the assumption that valid cues are to be found at the level of the reader's situation model. However, it posits that the valid cues are generated in response to disruptions that occur during the construction of a situation model. The present situation model approach forgoes the additional assumption that feedback from disruptions provides the cues that should be used to predict comprehension. Although this is one possible cue that a reader might use, it might not actually predict comprehension all that well. Disruptions resulting from construction and integration of representations may only indicate how difficult the process of comprehension was, which may not provide a basis to predict the degree of comprehension ultimately achieved, as initial disruptions might improve comprehension by prompting increased focus, attention, and effort by the reader. Further, for some of the findings cited above (the delayed keyword and summarization results) the interventions occurred well after any situation-model construction was complete. Thus, on the whole, these findings support the more general situation model approach, which posits that accurate metacomprehension depends on cues produced by accessing and utilizing one's situation-model.

Future work needs to continue to pursue the situation model approach as well as the demands of monitoring approach to improving metacomprehension accuracy. It will be important to fully understand what contexts may allow access to correct cues for predicting

comprehension as well as what contexts may support access to correct cues specifically in low ability learners.

More work is also needed to establish the relation between metacomprehension accuracy, self-regulation and learning. Thiede et al. (2003) showed that accuracy influenced the effectiveness of regulation and in turn overall reading comprehension. More recently Thomas and McDaniel (2007) have also provided some evidence that conditions that lead to better monitoring accuracy can also lead to better study choices and learning. However, these effects need to be replicated and extended to populations other than typical college students. It is not clear whether younger students or less-skilled readers can use their monitoring to guide regulation. Thus, future research should demonstrate how generally metacomprehension accuracy affects regulation of study and comprehension.

Author Notes

This research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305H030170 and R305B070018 to Keith Thiede, Thomas D. Griffin and Jennifer Wiley. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

- Anderson, M.C.M. & Thiede, K.W. (2008). Why Do Delayed Summaries Improve Metacomprehension Accuracy? *Acta Psychologica*, *128*, 110-118.
- Baker, J. M. & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin and Review*, *13*, 60-65.
- Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*, 610-632.
- Butler, D.L. & Winne, P.H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*, 245-281.
- Chi, M. T. H., DeLeeuw, N., Chiu, M., & Lavancher, C. (1994). Elicit self-explanation improves understanding. *Cognitive Science*, *18*, 439-477.
- Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, *12*(5), 769-86.
- Dunlosky, J., Baker, J. M., & Rawson, K. A. (2006). Does aging influence people's metacomprehension? Effects of processing ease on judgments of text learning. *Psychology and Aging*, *21*, 390-400.
- Dunlosky, J. & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*, 228-232.
- Dunlosky, J. & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*, 374-380.

- Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language*, 36, 34-49.
- Dunlosky, J., & Rawson, K. A. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes*, 40, 37-55.
- Dunlosky, J., Rawson, K., & Hacker, D. J. (2002). Metacomprehension of science text: Investigating the levels-of-disruption hypothesis. In J. Otero, J. A. Leon, and A.C. Graesser (Eds.). *The Psychology of Scientific Text Comprehension* (pp. 255-279). LEA, Mahwah, NJ.
- Dunlosky, J., Rawson, K. A., & McDonald (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. J. Perfect & B. L. Schwartz (Eds.) *Applied Metacognition* (pp. 68-92). New York, NY: Cambridge University Press.
- 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.
- Fischer, P. M. & Mandl, H. (1984). Learner, text variables, and control of text comprehension and recall. In H. Mandl, N. L. Stein, & T. Trabasso (Eds.) *Learning and Comprehension of Text* (pp. 213-254). London: Lawrence Erlbaum Associates.
- Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 702-718.

- Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, *116*, 119-136.
- 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.
- Hemmerich, J. & Wiley, J. (2002) Do argumentation tasks promote conceptual change about volcanoes. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Jee, B., Wiley, J & Griffin, T. D. (2006) Expertise and the illusion of comprehension. *Proceedings of the Annual Conference of the Cognitive Science Society*.
- Kelemen, W. L., & Weaver, C. A. (1997). Enhanced memory at delays: Why do judgments of learning improve over time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1394-1409.
- King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology*, *93*, 329-343.
- Kintsch, W. (1994). Learning from text. *American Psychologist*, *49*, 294-303.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.
- Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, *29*, 133-159.
- Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*, 609-639.

- 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.
- Lefevre, N. & Lories, G. (2004) Text cohesion and metacomprehension: Immediate and delayed judgments. *Memory & Cognition*, *32*, 1238-1254.
- Lin, L., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology*, *22*, 111-128.
- Lin, L. & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, *23*, 345-391.
- Lin, L., Zabrucky, K. M., & Moore, D. (2002) Effects of text difficulty and adults' age on relative calibration of comprehension. *American Journal of Psychology*, *115*, 187-198.
- Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *10*, 756-766.
- Maki, R. H. (1995). Accuracy of metacomprehension judgments for questions of varying importance levels. *American Journal of Psychology*, *108*, 327-344.
- Maki, R. H. (1998a). Test predictions over text material. In Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). *Metacognition in Educational Theory and Practice*. (pp. 117-144). Hillsdale, NJ: LEA.
- Maki, R. H. (1998b). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition*, *26*, 959-964.
- Maki, R. H. (1998c). Metacomprehension of text: Influence of absolute confidence level on bias and accuracy. In D. L. Medin, (Ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory* (pp. 223-248). San Diego, CA: Academic Press.

- 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*, 609-616.
- Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin and Review*, *1*, 126-129.
- Maki, R. H., & Serra, M. (1992a) The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 116-126.
- Maki, R. H., & Serra, M. (1992b). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology*, *84*, 200-210.
- Maki, R. H., & Serra, M. (1992b) Maki, R. H., Shields, M., Wheeler, A. E. & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology*, *97*, 723-731.
- McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*, 1-30.
- McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1-43.
- Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, *131*, 349-363.
- 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*, 267-270.

- Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, 26, pp. 125-141. New York: Academic Press, Inc.
- Nesbit, J. C. & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, 76, 413-448.
- Olin, J. T., & Zelinski, E. M. (1997). Age differences in calibration of comprehension. *Educational Gerontology*, 23, 67-77.
- Pierce, B. H. & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition*, 29, 62-67.,
- Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 69-80.
- Rawson, K., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*. 559-579.
- Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology*, 55, 505-524.
- Rawson, K., & Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, 28, 1004-1010.
- Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1998). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, 6, 132-137.
- Shaughnessy, J. J. & Zechmeister, E. B. (1992). Memory monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society*, 30, 125-128.

- Stensvold, M. S. & Wilson, J. T. (1990). The interaction of verbal ability with concept mapping in learning from a chemistry laboratory activity. *Science Education, 74*, 473-480.
- Thiede, K.W. (1999). The importance of accurate monitoring and effective self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6*, 662-667.
- Thiede, K. W. & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*, 129-160.
- 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.
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1024-1037.
- 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 & Cognition, 31*, 1267-1280.
- Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M.C.M. (in press). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*.
- Thomas, A. K. & McDaniel, M. A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition, 35*, 668-678.
- Vesonder, G. T. & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language, 24*, 363-376.

- Weinstein, C.E. & Mayer, R. E. (1986). The teaching of learning strategies. In M. C. Wittrock (Ed.), *Handbook on Research in Teaching* (3rd ed., pp. 315-327). New York: Macmillan.
- Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 214-222.
- Weaver, C. A., III, & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition*, 23, 12-22.
- Weaver, C. A., Bryant, D. S. & Burns, K. D. (1995). Comprehension monitoring: Extensions of the Kintsch and van Dijk model. In C. A. Weaver, Mannes, S., & Fletcher, C. (Ed.), *Discourse comprehension. Essays in honour of Walter Kintsch* (pp. 177-193). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Weaver, C. A. & Kelemen, W. L. (1997). Judgments of learning at delays: Shifts in response patterns or increased metamemory accuracy? *Psychological Science*, 8, 318-321.
- Wiley, J. (2001) Supporting understanding through task and browser design. *Proceedings of the Twenty-third annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Wiley, J., Griffin, T. & Thiede, K.W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology*, 132, 408-428.
- Wiley, J., Griffin, T. & Thiede, K.W. (2008). To understand your understanding you must understand what understanding means. *Proceedings of Annual Conference of the Cognitive Science Society*.

Wiley, J. & Voss, J. F. (1999) Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 301-311.

Winne, P. H. , & Perry, N. E. (2000). Measuring self-regulated learning. In M.

Boekaerts and P. Pintrich (Eds.), *Handbook of Self-Regulation*, pp. 531-566.

New York: Academic Press, Inc.

Table 1. *A Review of the Literature on Metacomprehension Accuracy*

Experimenters	Texts	Judgment	Test	N	Condition	Predict G	Postdict G	Predict G Effect Size
Anderson & Thiede, 2007	5 (~800 wds)	JOC 6pt	10 MC (inf)	87	No Summary	.20		.51 (summary conds)
	5 (~800 wds)	JOC 6pt	10 MC (inf)	(87)	Immediate Summary	.31		
	5 (~800 wds)	JOC 6pt	10 MC (inf)	(87)	Delayed Summary	.64		
Baker & Dunlosky, 2006 exp 1	8 (~400 wds)	PP 0-100	8 MC (4 det, 4 inf)	30	Immediate Judgement	.06		
				32	Delayed Judgement	.02		
exp 2	8 (~400 wds)	PP 0-100	8 MC (4 det, 4 inf)	33	Recall before Delayed Judgement	.10		
				31	Recall after Delayed Judgement	.18		
Dunlosky, Baker, & Rawson , 2006	8 (~400 wds)	PP 0-100	8 MC (4 det, 4 inf)	33	Younger Adults	.16		
				22	Older Adults	.23		
Dunlosky & Rawson, 2005	6 (~400 wds)	PP 0-100	6 MC (3 det, 3 inf)	39	Read once	.29		.256 (single vs immed)
				38	Immediate re-reading	.53		
				36	Delayed re-reading	.22		
Dunlosky, Rawson & , Hacker, 2002	6 (~400 wds)	PP 0-100	not stated?	38	No advance organizer	.36		
				35	Advance organizer	.42		
Dunlosky, Rawson & McDonald, 2002, exp 2	6 (~300 wds)	PP 0-100	4 cued recall	38	Global prediction	.37		
				(38)	Term-specific prediction	.38	.72	
Dunlosky, Rawson, & Middleton, 2005, exp. 1	6 (~300 wds)	PP 0-100	4 cued recall	26	Global Immed Judgement	.52		
				(26)	Term Immed Judgement	.57		
				27	Global Delayed Judgement	.05		
				(27)	Term Delayed Judgement	.57		
				24	Recall before Global Immed	.41		
				(24)	Recall before Term Immed	.73		
				26	Recall before Global Delayed	.52		
				(26)	Recall before Term Delayed	.64		
exp 2	6 (~300 wds)	PP 0-100	4 cued recall	39	Global Immed Judgement	.23		
				39	Term Immed Judgement	.61		
				40	Scored Recall before Global	.34		
				40	Scored Recall before Term	.57		
Griffin, Wiley, & Thiede, 2008, exp. 1	5 (~800 wds)	JOC 7-pt	~18 TF inf	36	Read once	.27		.337
				36	Reread	.45		
exp. 2	5 (~800 wds)	JOC 7-pt	~18 TF inf	30	Read once	.22		.212 (read vs. reread)
				30	Reread	.35		
				30	Reread and Self-explain	.67		
Jee, Wiley & Griffin, 2006	5 (~400 wds)	PP 6 pt	5 short answer	98	Multiple Texts on Baseball Topics	.20		
Lefevre & Lories, 2004 exp 1	6 (~550 wds)	JOC 0-20 pt	8 short answer	56	Immediate/paragraph judgements	.36		
				(56)	Delayed/global judgments	.43		
exp 2	6 (~550 wds)	JOC 0-20 pt	8 short answer	68	Immediate/paragraph judgements	.27		
				(68)	Delayed/global judgments	.38		
Lin, Moore & Zabucky, 2001	12 (~200 wds)	CONF 7-pt	4 TF inf	60	Read once	.14	.28	
Lin, Zabucky, & Moore, 2002	2 exp/2 narr	CONF 6-pt	16 MC (8 det/them)	120*	Easy	.19		.398 *(60 old/60 young)
				(120)	Standard	.47		
				(120)	Difficult	-.09		
Maki, 1995	12 (~200 wds)	PP 6-pt	3 MC (det)	18	Immed reread, Important ideas	.15	.56	

exp 1			3 MC (det)	18	Immed reread, Unimportant ideas	.01	.70	
			3 MC (inf)	18	Immed reread, Higher ideas	.27	.22	
exp 2	12 (~200 wds)	PP 6-pt	3 MC (det)	18	RSVP, Important ideas	.00	.22	
			3 MC (det)	18	RSVP, Unimportant ideas	.43	.47	
			3 MC (inf)	18	RSVP, Higher ideas	.36	.34	
Maki, 1998, exp 1	6 (~400 wds)	PP 6-pt	10 MC	16	Low Expect, Poorly anchor	.34	.47	
				16	Low Expect, No anchor	.03	.45	
				16	Low Expect, Well anchor	.04	.36	
				16	High Expect, Poorly anchor	.27	.47	
				16	High Expect, No anchor	.18	.33	
				16	High Expect, Well anchor	.47	.63	
exp 2	6 (~400 wds)	PP 6-pt	10 MC	23	Low Expect, Poorly anchor	.13	.30	
				24	Low Expect, Well anchor	.14	.19	
				23	No Expect, Poorly anchor	.23	.36	
				24	No Expect, Well anchor	.17	.42	
				23	High Expect, Poorly anchor	.12	.39	
				24	High Expect, Well anchor	.13	.57	
Maki, 1998 exp 1	4 (~400 wds)	PP	16 MC (det)	24	Multiple Texts/Topics	.28		
exp 2	1 (~1600 wds)	PP	64 MC (det)	24	Single Text/Topic	.14		
	6 (~400 wds)	PP	5 MC	32	Familiarity predicts Performance	.38		
	6 (~400 wds)	PP	5 MC	32	Familiarity Does not Predict	.37		
Maki, 1998	6 (~400 wds)	PP 6-pt	5 MC	24	Immediate Rating and Test	.47	.62	.323
				24	Immed Rating, Delayed Test	.05	.40	
				24	Delayed Rating, Immed Test	.20	.49	
				24	Delayed Rating, Test	.02	.46	
Maki, Foley, Kajer, Thompson, & Wilert, 1990, exp. 1	1 (21-27 para)	EOC 6-pt	2 cued recall per p	40	EOC-intact text	.26	.45	.238 (intact vs. deleted)
		CONF 6-pt		40	EOC-deleted letters	.13	.55	
				(40)	CONF-intact text	.16	.56	.334 (intact vs. deleted)
				(40)	CONF-deleted letters	.31	.59	
exp. 2	1 (~14-17 para)	CONF 6-pt	2 cued recall per p	40	intact text	.18	.52	.321
				40	deleted letters	.37	.53	
Maki, Jonas & Kallod, 1994	1 (~1700 wd)	PP 6-pt	4 MC	75	read once	.11	.55	
Maki & Serra, 1992a exp. 1	12 (~200 wds)	PP rank titles	4 MC	36	practice test	.22	.35	.146
exp. 2	12 (~200 wds)	PP rank titles	3 MC	36	no practice test	.23	.32	
				18	identical to practice	.32	.24	.230
				18	stem from practice	.23	.24	
				18	similar to practice	.10	.45	
				18	no practice	.27	.40	
exp. 3	12 (~200 wds)	PP rank titles	3 MC	18	no feedback on similar practice test	.29	.37	.360
				18	feedback on similar practice test	.06	.21	
Maki & Serra, 1992b exp. 1	12 (~200 wds)	EOC rank titles	4 MC	48	comprehension judgement	.17	.28	
exp 2	12 (~200 wds)	PP rank titles	4 MC	48	performance judgement	.31	.40	
		EOC rank titles	4 MC	24	comprehension judgement	.14	.34	
		PP rank titles	4 MC	24	performance judgement	.40	.41	
exp 3	12 (~200 wds)	EOC rank titles	6 MC	18	comprehension judgement	.11	.25	
		PP rank titles	6 MC	18	performance judgement	.25	.31	

Maki, Shields, Wheeler, & Zacchilli, 2005	6 (~400 wds)	PP 7 pt	6 MC (3 det, 3 inf)	52	Easy Texts	.20	.36	Not reported
	6 (~400 wds)	PP 7 pt	6 MC (3 det, 3 inf)	55	Mixed Set	.31	.42	
	6 (~400 wds)	PP 7 pt	6 MC (3 det, 3 inf)	51	Hard Texts	.36	.57	
Miesner & Maki, 2007	6 (~400 wds)	PP 0-100	5 MC	38	Low Test Anxiety	.20	.19	
				42	High Test Anxiety	.58	.46	
Olin & Zelinsky, 1997	4 (9 para)	JOC 6 pt	36 TF sent recog	51	Younger	.11		
				52	Older	.13		
Pierce & Smith, 2001, exp 2	3 (~500 wds)	PP 5 pt	16 MC (8 det/them)	47	read once	.28	.53	
Rawson & Dunlosky, 2007	6 (~300 wds)	PP 0-100	4 cued recall	26	Term recall/score without answer	.68		
				30	Term recall/score with answer	.52		
				(26)	Global/score without answer	.21		
				(30)	Global/score with answer	.36		
Rawson, Dunlosky, & Thiede, 2000, exp 1	6 (~400 wds)	PP 0-100	6 MC (inf?)	40	read once	.24	.50	.298
				40	rereading	.57	.47	
exp. 2	6 (~200 wds)	JOC 7-pt	6 MC (inf?)	30	read once	.19		.413
				30	rereading	.55		
Thiede & Anderson, 2003, exp. 1	6 (~200 wds)	JOC 7-pt	6 MC (thematic?)	25	No Summary	.28		.308
				25	Immediate Summary	.24		
				25	Delayed Summary	.60		
exp. 2	6 (~1300 wds)	JOC 7-pt	12-MC (6 det, 6 inf)	30	No Summary	.28		.262
				30	Immediate Summary	.35		
				30	Delayed Summary	.61		
Thiede, Anderson, & Therriault, 2003	6 (~1300 wds)	JOC 7-pt	6-MC (3 det, 3 inf)	22	No keyword generation	.38		.351
				22	Immediate keyword generation	.29		
				22	Delayed keyword generation	.70		
Thiede, Dunlosky, Griffin, & Wiley, 2005, exp. 1	6 (~1300 wds)	JOC 7-pt	12-MC (6 det, 6 inf)	33	Delayed Keyword and Judgement	.52		.315 (delay vs. immediate)
				33	Delayed Keyword, Immed Judge	.55		Not significant (lag vs. no-lag)
				33	Immediate Keyword, Delay Judge	.27		
				33	Immediate Keyword and Judge	.29		
exp. 2	6 (~1300 wds)	JOC 7-pt	6 MC (6 inf)	25	Delayed, Successive Keyword	.57		.354 (delay vs. immediate)
				25	Delayed Spaced Keyword	.55		
				25	Immediate Spaced Keyword	.16		
				25	Short delay, Spaced Keyword	.39		
exp. 3	6 (~1300 wds)	JOC 7-pt	6 MC (6 inf)	27	Immediate-think-about-text	.25		Not significant
				27	Delayed-think-about-text	.27		
exp. 4	6 (~1300 wds)	JOC 7-pt	6 MC (6 inf)	33	Delayed Keyword generation	.52		.338
				33	Delayed keyword reading	.27		
Thiede, Griffin, Wiley & Anderson, in press, exp 1	5 (~800 wds)	JOC 7-pt	10 MC (inf)	74	Normal reader, No Summary	.21		.276 (normal vs. at-risk)
				(74)	Normal reader, Immed Summary	.29		.196 (summary conds.)
				(74)	Normal reader, Delayed Summary	.63		
				32	At risk reader, No Summary	.08		
				(32)	At risk reader, Immed Summary	.17		
				(32)	At risk reader, Delayed Summary	.48		
exp 2	5 (~300 wds)	JOC 7-pt	5 MC (inf)	21	At risk, Immediate Judgement	.30		.424 (conds.)
				(21)	At risk, Delayed Judgement	.33		
				(21)	At risk, Concept Map Generation	.66		

Thomas & McDaniel, 2006, exp. 1	6 (~300 wds)	PP 0-100 scale	6 cued recall (det)	39	letter reinsertion, detail	.41	.334 (task)
			6 cued recall (them)	39	letter reinsertion, conceptual	-.33	.489 (task x test)
			6 cued recall (det)	39	sentence sorting, detail	-.06	
			6 cued recall (them)	39	sentence sorting, conceptual	.59	
			6 cued recall (det)	39	reading, detail	.27	
exp. 2	6 (~300 wds)	PP 0-100 scale	6 cued recall (them)	39	reading, conceptual	.11	
			6 cued recall (det)	25	letter reinsertion, detail	.62	.494 (task x test)
			6 cued recall (them)	25	letter reinsertion, conceptual	-.13	Not reported (task)
			6 cued recall (det)	25	sentence sorting, detail	.06	
			6 cued recall (them)	25	sentence sorting, conceptual	.55	
Weaver, 1990, exp. 1	16 (~200 wds)	PP 6-pt	1 TF (inf)	21	one item test	.09	Not reported
exp. 2	15 (~200 wds)	PP 6-pt	1 TF (inf)	18	one item test	.02	Not reported
			2 TF (inf)	16	two item test	.28	
			4 TF (inf)	19	four item test	.40	
Weaver & Bryant, 1995	4 (~500 wds)	JOC 6-pt	8 MC Them	49	Thematic	.22	.268
exp. 1			8 MC Detail	(49)	Detailed	.39	
exp. 2	2 narr/2 exp	JOC 6-pt	8 MC Them	30	Easy-Thematic	.20	.267
			8 MC Detail	(30)	Easy-Detailed	.38	
			8 MC Them	31	Standard-Thematic	.75	
			8 MC Detail	(31)	Standard-Detailed	.63	
			8 MC Them	31	Difficult-Thematic	.33	
exp. 3	2 narr/2 exp	JOC 6-pt	8 MC Detail	(31)	Difficult-Detailed	.30	
			8 MC Them	30	Easy-Thematic	.15	.168
			8 MC Detail	(30)	Easy-Detailed	.37	
			8 MC Them	31	Standard-Thematic	.72	
			8 MC Detail	(31)	Standard-Detailed	.67	
Wiley, Griffin & Thiede, 2008, exp 1	6 (~800 wds)	PP 6-pt	8 MC Them	31	Difficult-Thematic	-.04	
			8 MC Detail	(31)	Difficult-Detailed	.45	
			5 MC Detail	36	No Test Expectancy	.23	
			5 MC Inf	(36)	No Test Expectancy	.22	
			5 MC Detail	36	Expect Memory Test	.38	
			5 MC Inf	(36)	Expect Memory Test	.18	
			5 MC Detail	36	Expect Inference Test	.17	
			5 MC Inf	(36)	Expect Inference Test	.36	

exp 2	6 (~800 wds)	PP 6-pt	5 MC Detail	36	No Expectancy	.34
			5 MC Inference	(36)	No Expectancy	.16
			5 MC Detail	36	Expect Inference Test	.28
			5 MC Inference	(36)	Expect Inference Test	.38
			5 MC Detail	36	Self Explain	.15
			5 MC Inference	(36)	Self Explain	.32
			5 MC Detail	36	Expect Inference, Self Explain	.20
			5 MC Inference	(36)	Expect Inference, Self Explain	.51

PP = performance prediction, e.g., “How well do you think you will be able to answer test questions over this material?” 0 (*definitely won't be able*), ...,100 (*definitely will be able*) or number of items

JOC = judgment of comprehension, e.g. “How well do you think you understand the text you just read?” 1 (not at all) to 7 (very well)

EOC = judgement of how easy it was to comprehend a text

CONF = judgement of confidence, e.g. “report your confidence that you will be able to use what you learned to answer test questions” 1 (very low) to 6 (very high)

exp. = experiment, MC = multiple choice, TF = true/false, det = detail questions, inf =inference questions, them = thematic questions

All texts are expository (exp) unless otherwise noted as narr = narrative

N in parenthesis indicates repeated measures design

G = gamma correlations