Improving Students’ Metacomprehension Accuracy

Reading assignments are still a primary means by which students learn in their courses. Often, students will have to complete readings for multiple topics within a limited amount of time. And, as course exams approach, they must regulate their time and efforts in what and how they study in order to optimize learning and test performance across topics and courses. Theoretically, optimal study regulation will depend on how accurately students can monitor and self-assess their comprehension of various pieces of text, in order to avoid wasting time by restudying mastered material or by continuing to employ ineffective strategies that are not leading to comprehension (Thiede & Dunlosky, 1999; Winne & Hadwin, 1998). This monitoring of ongoing learning processes has been referred to as metacognitive monitoring or, in the context of text comprehension, metacomprehension (Maki & Berry, 1984). How accurate learners are in monitoring and judging their progress is metacomprehension accuracy.

The earliest research to approach the issue of monitoring during text comprehension employed error-detection paradigms (e.g., Markman, 1977). Readers’ failure to notice or mention errors in texts has been claimed to indicate their failure to monitor the coherence of their mental representations of the text. However, there are many alternative explanations for readers’ failure to mention errors, such as a lack of comprehension itself or a presumption that the texts are error-free and that any confusion is their own fault (see Markman, 1979). Other research has attempted to use online measures of reading behaviors (e.g., reading slow-downs when encountering contradictions) to show error detection and thus implicate comprehension monitoring (Otero, 1998; Wiley & Myers, 2003), but many alternative explanations exist for slow-downs in reading times as well.

The earliest studies that attempted to directly assess the accuracy of monitoring during text comprehension using a judgment paradigm were initiated by Maki and Berry (1984), and Glenberg and Epstein (1985). The basic paradigm is outlined in

* The authors were supported by grants from the National Science Foundation (DUE 1535299) and the Institute of Education Sciences (R305A160008) during the preparation of this chapter. All opinions expressed herein are those of the authors and do not necessarily reflect those of the funding agencies. The authors thank Keith W. Thiede and Tricia A. Guerrero for their comments on this chapter.
Figure 24.1 (minus the effective manipulations shown in the shaded box). It involves participants reading multiple texts, providing predictive judgments of comprehension for each text, and answering test questions for each text. These researchers borrowed the paradigm from two decades of research on metamemory, where learners were asked to predict their future ability to recall/recognize general facts, presented pictures, or studied word pairs (e.g., Flavell, Freidrichs, & Hoyt, 1970; Hart, 1965; Nelson & Narens, 1980). These simpler learning materials were replaced with texts typically ranging from 200 to 1,000 words, and the monitoring judgments about memory were replaced with judgments of “comprehension,” “understanding,” or predictions of performance on a “test of comprehension.” The majority of research on metacomprehension of text has followed Nelson’s (1984) advice regarding measures of metamemory accuracy, by focusing primarily on relative accuracy computed via intra-individual correlations between an array of judgments and an array of test performances rather than difference scores between judgments and tests, whether signed or absolute.

Since these initial studies, there have now been over eighty experiments that have measured the relative accuracy of judgments of comprehension for texts. Across studies, methodological approaches have varied greatly, from the length of text that corresponds to each monitoring judgment to whether the tests focus on memory for the isolated ideas explicitly stated in the text versus comprehension of the implied logical and causal relations among various concepts. In this chapter, the former type of test is considered to reflect metamemory for text, whereas the latter is considered to reflect metacomprehension. As suggested by both theory and empirical findings, accurate metamemory and metacomprehension for text are distinct constructs, rely on different types of judgment cues, and can be impacted in opposite ways by a given EI Nino

Ice Ages

Volcanoes

Evolution

CO2 Cycle

Lightning

EI Nino

Lightning

CO2 Cycle

Evolution

Volcanoes

Ice Ages

Study passages on 6 different topics

Manipulations that improve metacomprehension

Delayed Keyword Generation

Delayed Summary Generation

Generating Explanations

Generating Concept Maps

How many questions out of 5 will you get right?

EI Nino 3

Lightning 6

CO2 Cycle 3

Evolution 5

Volcanoes 4

Ice Ages 4

Predict test scores on 6 different topics

Take 5-item tests on 6 different topics

Figure 24.1 Overview of the standard metacomprehension paradigm (minus the shaded box)

The shaded box shows manipulations that increase access to situation-model–based cues and improve metacomprehension accuracy.
manipulation. The main focus of this chapter is providing an overview of effective manipulations specifically for improving metacomprehension. In general, students’ ability to accurately monitor their own comprehension of text is quite poor. Several reviews have reported average levels of metacomprehension accuracy around 0.27 in baseline conditions without special instructions or activities (Dunlosky & Lipko, 2007; Maki, 1998a, 1998b; Thiede et al., 2009). However, a substantial body of evidence has begun to accumulate suggesting that there are several instructional conditions that can lead to significant improvements, such as those shown in the shaded box in Figure 24.1. Before turning to the research on these effective manipulations, it is important to situate that literature within a theoretical context and discuss relevant methodological concerns.

**Monitoring and Other Components of Metacognition**

Hacker (1998) identified two different concepts in the literature commonly referred to as “monitoring.” An approach used primarily by educational researchers uses “comprehension monitoring” in a broad sense to refer to any deliberate attention that learners pay to their learning goals, use of strategies, and learning progress. This approach generally utilizes assessments such as self-report scales of strategy knowledge and strategy use, rather than focusing on accurate monitoring of ongoing learning (for a review, see Dinsmore, Alexander, & Loughlin, 2008). The second approach, used primarily by cognitive psychologists, focuses on learners’ monitoring of ongoing learning. Monitoring is measured by having students make explicit judgments of their current level of learning or understanding. Using this approach, comprehension monitoring accuracy can be computed by comparing judgments to performance on objective tests of comprehension. Identifying the conditions that improve performance on these metacomprehension accuracy measures is the focus of this chapter.

Monitoring progress of ongoing learning was central to Flavell’s (1979) original conception of metacognition and he distinguished it from the types of more general meta-knowledge that are assessed by self-reported study strategies. Metacognition entails two levels of processing that are reflected in the predominant models of metacognition and self-regulated learning (SRL; Griffin, Wiley, & Salas, 2013; Metcalfe, 2002; Nelson & Narens, 1990; Pintrich, Wolters, & Baxter, 2000; Thiede & Dunlosky, 1999; Winne & Hadwin, 1998; Zimmerman, 2002). As shown in the shaded portion at the bottom of Figure 24.2, object-level processing entails the application of a priori knowledge and assumptions about particular tasks, strategies, or a learner’s abilities that can be used to implement cognitive actions to increase learning. The upper unshaded area is meta-level processing that entails subjective experiences of internal states that occur as a result of the cognitive actions and reflect how learning is progressing. Only by monitoring these experiences during a learning episode can one be aware of actual variations in learning that arise from interactions among the specific texts or learning material, contextual factors, and how well the reader actually executes any learning
strategies. Such monitoring allows the learner to make adjustments and regulate learning beyond what is afforded by limited, and often inaccurate, a priori assumptions and expectations about learning. This feedback loop of actively monitoring subjective experiences during learning is what Flavell meant by “monitoring” and makes metacognition qualitatively distinct from the cognitive processes or operations that learners engage in at the “object” level (Fischer & Mandl, 1984; Griffith et al., 2013; Nelson & Narens, 1990).

The focus of research on monitoring accuracy has been to identify effective metacognitive strategies, which, as Flavell (1979) argued, are strategies specifically for aiding the accurate monitoring of learning rather than more general cognitive/study strategies that directly aid learning itself. The distinction lies in the learner’s purpose for using a strategy. The same activity (e.g., practice tests or self-explaining what was just read) could be employed as a cognitive strategy to improve retention or comprehension or, alternatively, it could be used as a metacognitive strategy to generate meta-experiences to serve as cues of how well one already comprehends the material. Learners know and can apply cognitive strategies that impact object-level processing and directly affect learning outcomes without involving meta-level processes. Nearly all models of SRL have at least an implicit recognition that the mere application of study strategies does not inherently entail meta-level processing or monitoring of ongoing learning (e.g., Hacker, Keener, & Kircher, 2009; Nelson & Narens, 1990; Pintrich et al., 2000; Winne & Hadwin, 1998). Instead, what is critical is that students are using a strategy specifically to actively reflect on their learning progress.

**How Are Metacognitive Judgments Made?**

The prevailing theoretical framework that guides research on monitoring accuracy is Koriat’s (1997) cue-utilization theory, which presumes that learners cannot directly observe their own level of understanding and must make use of
indirect cues to infer it. Koriat discussed two classes of cues from which learners
draw inferences that form the basis of their judgments and performance predictions.
One class is comprised of cues that are tied to learners’ internal online subjective
experiences that result from their cognitive processing in the specific situation.
Koriat labeled these mnemonic cues because his focus was on monitoring during
memorization tasks. These cues include the subjective sense of ease or fluency
during learning (Benjamin & Bjork, 1996; Dunlosky & Nelson, 1992). The other
kinds of cues are tied to objective features of the learning situation. These cues can be
either intrinsic to the materials and task demands (e.g., relatedness of word pairs,
memory of details versus conceptual application) or extrinsic to the task or stimuli
but instead related to the context (e.g., how many times items were studied or what
strategy was used). Griffin, Jee, and Wiley (2009) pointed out that these intrinsic and
extrinsic cues depend on the application of a priori knowledge or assumptions,
whereas the mnemonic cues are meta-experiences tied to actually constructing
a particular mental representation during a learning episode. Thus, the explicit
judgments of learning (JOLs) that people are asked to make can either be inferred
based in actual monitoring of meta-experiences or be inferred based on general
heuristic knowledge or assumptions about learning.

General heuristic knowledge may sometimes be useful in predicting aggregate
levels of learning because such knowledge is informed by the information one has
acquired about oneself and various factors related to learning from prior experience.
If one has generally done better on tests than one’s peers, then knowledge of this
could accurately predict doing better than one’s peers on average on any future test.
In fact, such heuristic knowledge could be used to predict other people’s under-
standing as well as one’s own. However, such knowledge will not aid in the
prediction of within-person variability in learning because it fails to incorporate
the countless and often highly context-dependent factors that impact comprehen-
sion of a particular text at a particular moment. Predicting context-dependent learning
that varies within-person across topics requires attention to those subjective, online
meta-experiences at the center of the metacognitive monitoring construct. This
general point leads to two more specific arguments. The first is that it is preferable
to have a measure of monitoring accuracy that is sensitive to these online experi-
ences. The second is that it is preferable if that measure also is sensitive to diagnostic
cues that predict actual comprehension. These two points are discussed further in the
following sections.

### Differences Between Measures of Monitoring Accuracy:
Relative Versus Absolute Measures

Research on monitoring accuracy relies on having learners make explicit
judgments about their level of learning, understanding, or comprehension (JOLs,
JOUs, JOCs), or predictions of performance (POPs) on upcoming tests. The
accuracy of these judgments is determined by comparing them with actual
performance on subsequent tests assumed to be objective measures of actual
comprehension. Various measures of metacomprehension compare predictions with actual performance but they do so in different ways. Three independent measures of judgment–performance relations have been employed by metacomprehension researchers: **absolute accuracy**, **confidence bias**, and **relative accuracy** (Griffin et al., 2009; Maki 1998a). **Absolute accuracy** is the mean of the absolute (or mean squared) deviations between judged and actual performance. For absolute accuracy measures, lower values indicate better accuracy. A related outcome measure, **confidence bias**, concerns the directionality of these deviations, computed as the mean of the signed difference between each judgment and corresponding test score. This measure, sometimes referred to as overconfidence (for positive values) or underconfidence (for negative values), does not reflect the level of accuracy in terms of the number or magnitude of judgment errors but rather in terms of whether those judgment errors are systematically biased in one direction (Yates, 1990). Both absolute accuracy and confidence bias are statistically dependent on marginal mean performance levels and this allows for nonmetacognitive influences on these accuracy measures (for a longer discussion, see Nelson, 1984).

Therefore, a third measure, **relative accuracy**, has become the main measure of interest for most metacomprehension research. Relative accuracy quantifies an individual’s accuracy in predicting performance on one text relative to other texts in terms of an intra-individual correlation between the individual’s performance predictions and the individual’s test scores (e.g., Gamma or Pearson). Nelson (1984) strongly recommended computing relative accuracy using intra-individual correlations over measures of absolute accuracy because the latter are so directly dependent on aggregate levels of test performance. Two people who engage in identical metacognitive processes will wind up with drastically different absolute accuracy scores simply because one did worse on the tests. Similarly, the overall difficulty of the tests can increase or decrease absolute accuracy.

To appreciate the differences between these measures, consider Figure 24.3 that illustrates two sets of hypothetical predictive judgments and two sets of performance on comprehension tests for texts on six different topics. For the purposes of this illustration, the topics have been arranged in order from the text that was understood the best to the text that was understood the least, for a set of easy tests and a set of hard tests. Relative accuracy is the intra-individual correlation between the array that results from an individual learner’s judgments and the array that results from that individual’s actual performance on the corresponding tests. If the reader makes Judgment A on each topic, then it would result in perfect relative accuracy (both with the easy and hard tests) because the judgments vary in parallel with variations in test performance. In contrast, if the reader makes Judgment B on each topic, it would result in weak relative accuracy (at the typically observed 0.28) because these judgments sometimes vary in the opposite direction relative to test performance. Importantly, the general difficulty level of the tests does not impact the relative accuracy scores. However, the general difficulty level of the tests makes a large difference in absolute accuracy. When absolute accuracy is computed as the mean of the absolute differences between each judgment–performance pair, lower scores mean less error and greater accuracy. When the tests are generally easy, Judgment
A would lead to much better absolute accuracy than Judgment B (20 versus 40.17). Whereas, for difficult tests, Judgment A would have far worse accuracy than Judgment B (30 versus 10.33).

Although it is pragmatically useful for learners to know whether they have mastered material in an absolute sense (Dunlosky & Rawson, 2012), measures of absolute accuracy have less utility for advancing theories of monitoring because they confound effects due to object-level processes with meta-level processes. In some cases, this problem can be reduced by using comprehension performance measures as statistical controls when analyzing effects on absolute metacomprehension accuracy (Griffin et al., 2009). In contrast, relative accuracy provides a measure that is orthogonal to test performance and average judgment magnitude. This means it is also orthogonal to absolute accuracy so they are not measures of the same psychological construct. The differences between the constructs of absolute and relative accuracy are important theoretically as relative accuracy is the more direct reflection of how well a learner is engaging in effective monitoring of meta-experiences during learning. Maki and colleagues (2005) pointed out that “participants have to use their experience with the specific task in order to produce accurate relative judgments across parts of that task” (p. 729, emphasis added). The low levels of relative accuracy that are generally obtained in studies of metacomprehension (Dunlosky & Lipko, 2007; Maki, 1998a, 1998b; Thiede et al., 2009) may result from individuals’ inability to distinguish their level of learning from specific texts from their assumed ability to learn from text in general. In contrast, heuristic knowledge about one’s general ability can produce judgments that are accurate in an absolute sense. As a byproduct of how they are computed, absolute accuracy measures are sensitive

Figure 24.3 Relative versus absolute accuracy and test difficulty
to anything that impacts average judgment magnitude, such as a priori heuristic assumptions. In contrast, high relative accuracy can only be achieved by relying on cues that reflect the context-dependent variations in how well one is actually comprehending information from text to text and from moment to moment. In other words, differences in relative accuracy will be more reflective of how well learners are actually monitoring meta-experiences created while processing the material.

Several studies have indirectly illustrated this point by showing the greater sensitivity of absolute accuracy measures to differences in learners’ heuristic assumptions. Griffin and colleagues (2009) showed that having more general domain knowledge and self-reported prior domain familiarity predicted higher judgments and greater absolute accuracy but not relative accuracy. This held after controlling for knowledge effects on test performance itself and even when participants only saw text titles and did not even read the texts before making their judgments and taking the tests. In addition, an index of how much each participant based their judgments on prior familiarity was correlated with absolute accuracy scores. In contrast, prior general knowledge and familiarity had no relation with relative accuracy. Thus, when readers used a priori heuristic knowledge unrelated to meta-experiences during text processing to make judgments, this led to better absolute accuracy but not better relative accuracy.

Other studies have suggested that absolute accuracy is more sensitive to heuristic cues, although this point was not made explicitly by the authors of those studies. Generalized a priori expectations about how hard the tests will be are a form of general heuristic knowledge independent of monitored meta-experiences. Maki (1998a) manipulated learners’ prior heuristic knowledge about test difficulty by telling them that the “typical” score was either high or low. This heuristic knowledge had a large impact on learners’ judgment magnitudes and the group that was told that the tests were easy had much higher absolute accuracy in both experiments ($d = 0.74$ and $0.99$), whereas this knowledge had only a small effect on relative accuracy in one experiment ($d = 0.29$) and no effect in the second experiment. Zhao and Linderholm (2011) found the same effect, except in the reverse direction, such that the group given the expectancy of difficult tests had the best absolute accuracy. In Maki (1998a) the tests happened to be easy (75 percent correct), whereas the tests were more difficult (50 percent correct) in Zhao and Linderholm (2011). This is an empirical example of the problem illustrated in Figure 24.3 in which the same manipulation is used in two studies, and may have the same impact on metacognitive processes, and yet produces opposite results in terms of absolute accuracy depending on how difficult or easy the researchers chose to make the tests (for a conceptually similar argument, see Connor, Dunlosky, & Hertzog, 1997).

Besides heuristics about tests, there are also heuristics that learners hold about themselves, the learning materials, and different media that help them learn. Kwon and Linderholm (2014) found that students’ a priori heuristic beliefs about their general reading skill predicted their overall judgment magnitude for specific texts ($r = 0.44$) even better than their actual reading skill ($r = 0.29$), and predicted absolute accuracy and confidence bias. Ackerman and Goldsmith (2011) found that when
learners were given a set amount of study time for each text, there was a bias for making higher judgments when the texts were read on computer screens versus on paper, despite no effect of text presentation on actual test performance. On-screen reading produced higher overconfidence and worse absolute accuracy but there was no impact on relative accuracy.

Ackerman and Goldsmith’s (2011) findings suggest that readers apply a general heuristic about learning with technology, which is related to the heuristic about multimedia learning suggested by Serra and Dunlosky (2010). In their experiments, Serra and Dunlosky collected an initial set of predictive judgments prior to reading, when learners only knew they would read “a short science text about how lightning storms develop” in the text-only condition or “a short science text with diagram” in the multimedia conditions. Serra and Dunlosky found higher prereading judgments in the multimedia conditions than the text-only condition. Because readers had not yet seen the texts, these higher judgments could only be due to learners applying a general heuristic that multimedia helps learning. Further, reliance on this heuristic persisted into a second set of postreading judgments, even after learners read the texts and saw that the “diagrams” were merely decorative photos without conceptual relevance to how lightning storms develop. As shown in table 3 of their paper (p. 705), the test items were generally very difficult, with participants on average getting only 33 percent correct, whereas average postreading judgments of test performance were nearly twice as high (62 percent), indicating overconfidence in all but a few participants. Serra and Dunlosky did not report absolute accuracy but, given that almost all participants were overconfident in all conditions, absolute accuracy can be estimated from the mean performance and mean per-paragraph judgment scores reported in table 3 of their paper. The inflated judgments in the photos condition led to worse estimated absolute accuracy in the multimedia conditions compared to the text-only condition (37.3 vs. 29.7). However, the same heuristic also led to better absolute accuracy in a multimedia condition that had conceptually useful images. This multimedia condition produced inflated judgments similar to those seen in the nonuseful photo condition but also markedly increased test performance up to 43 percent correct. This resulted in an estimated absolute accuracy that was actually better (23.6) than that of the text-only condition. In contrast, the per-paragraph relative accuracy was unaffected by the multimedia heuristic.

In sum, absolute accuracy appears highly sensitive to variability in heuristic cues that are not tied to monitoring of meta-experiences during learning and varies depending on test difficulty. While better absolute accuracy may sometimes reflect more accurate heuristic assumptions held by readers, it can also reflect whether these assumptions happen to match the average difficulty level of the tests created by researchers or instructors.

The previously mentioned measure of confidence bias, the simple signed average of judgment−performance difference scores, shares many of the interpretation issues as absolute accuracy because it is also orthogonal to relative accuracy and dependent on mean test performance and mean judgments. However, when construed as a measure of monitoring accuracy, confidence bias is particularly problematic.
In contrast to the standard absolute accuracy measure, and as its name implies, this bias score does not reflect accuracy in terms of how much error was in a person’s judgments, but rather it reflects the extent to which a person’s errors were either random (thus cancelling each other out to produce a bias score close to zero) or systematically biased in one direction (to overestimate or underestimate performance). When comparing conditions in a study, mean bias scores do not reflect the magnitude of the difference between judgments and performance, but rather whether the people in a condition shared the same directional bias. Since the least amount of bias (a difference score of zero) falls in the middle of all possible bias scores, a difference between means may not even reflect more or less bias but merely that one subgroup was overconfident while another was equally underconfident. Bias and absolute accuracy can be unrelated or negatively related but will be positively related if the majority of a sample happens to share the same directional bias (such as when the tests are generally difficult, so performance is consistently below judgments). If researchers create especially hard tests to ensure variability on their measure, this is likely to produce general overconfidence. In sum, absolute accuracy, relative accuracy, and confidence bias are the three most common outcome measures reported in metacomprehension research and are sometimes treated as interchangeable despite being largely unrelated. These measures generally are not suited to test the same theories or hypotheses.

A final measure that is reported from time to time is an inter-person correlation, where a mean judgment and mean performance score is calculated for each person, then the two measures are correlated at the level of study condition. Readers of this literature need to be aware that there are instances where such “inter-individual” correlations are incorrectly reported as “intra-individual” correlations. Some signals that the correlations are inter-individual are when they are reported as correlation coefficients tested against zero. With relative accuracy intra-individual correlations, the actual Gamma or Pearson scores are only for an individual person so they would not be reported as correlations with confidence intervals. Instead, mean scores and variation indices (like standard deviations) would be reported for each study condition and tested against each other.

Pilegard and Mayer (2015) argue that an inter-person correlation reflects whether learners can predict their own performance relative to the performance of others. However, it is unclear what this construct could mean or how it could inform regulation of study behaviors. There is no statistical or conceptual overlap between these inter-individual correlations and relative accuracy as computed by intra-individual correlations. This inter-individual measure also does not have a straightforward connection to absolute accuracy. The inter-person correlation reflects whether the people within a given condition share the same direction and magnitude of confidence bias. Two conditions could be identical in absolute error but, if the errors in one condition are more systematically biased in the same direction, then the mean confidence bias will be greater and the inter-individual correlation will be stronger in that condition. This makes it unclear whether such as a measure has any theoretical or pragmatic utility.
Gamma versus Pearson as relative accuracy measures. Although Nelson (1984) suggested using intra-individual Goodman–Kruskal Gamma correlations as the basis for computing relative accuracy for studies on metamemory, other arguments can be made in favor of using Pearson to compute relative accuracy for studies on metacomprehension. The original recommendation for using Gamma was based in metamemory work where performance measures (memory tests) were dichotomous (recall or no recall). In such instances, it makes sense to compute a tally of hits and misses as an accuracy measure, which is what is involved in computing Gamma. Computing Gamma entails making a set of pairwise comparisons, tabulating the number of times that two judgments differ in the same direction (concordance/hit) versus opposite direction (discordance/miss) as the difference between the corresponding pair of test performances. Gamma attends only to the frequency of hits and misses, completely ignoring the large variance in how close those hits and misses actually are. The Pearson–Gamma tradeoff is one of making an assumption of linearity in the variance of the judgment scale versus completely discarding most of the variance information that nondichotomous scales yield.

In contrast to metamemory studies, most studies in the metacomprehension literature use comprehension tests with multiple items and are therefore collecting more continuous measures of performance. When comparing Pearson and Gamma even for dichotomous performance measures, Schwartz and Metcalf (1994) noted “a subject might be revealing something real and interesting . . . by giving values of .49, .50, and .51 for the three items in one list [rather than] .01, .50, and .99,” yet “gamma discards this information” (p. 105). Griffin, Wiley, and Thiede (2008) further noted that when judgments are predicting continuous comprehension scores derived from multi-item tests, then the case for Pearson is stronger due to the extreme amount of information that Gamma discards. Also, the linearity assumption is more sound when readers are judging the number of test items they will correctly answer rather than endorsing subjective abstractions like being “somewhat confident” in one’s answer. It is easy to find fault with any correlational method based on only six judgment–test pairs per participant, as is common in studies on text metacomprehension. However, there is no easy solution to this problem given the goal of exploring the monitoring of deeper comprehension processes, which requires using texts of sufficient length and complexity to support the construction of explanatory models and a single judgment for each text that reflects the reader’s monitoring of the quality and completeness of their representation of each explanation as a whole. This generally makes collecting a much larger number of judgment–performance pairs untenable in metacomprehension experiments.

Benjamin and Diaz (2008, p. 78) provide evidence that a variant of signal-detection measures ($D_a$) is preferable to Gamma, in part, because it does not “discard vast amounts of information” that the pairwise approach of Gamma does. According to their approach, $D_a$ gives consideration to the number of judgment levels between any two judgments (however, for a $D_a$ measure based on the same dichotomous hits and misses categorizations as Gamma, see Masson & Rotello, 2009). Although Benjamin and Diaz (2008) also critique Pearson for its linearity assumption, the simulations and conclusions resulting from their approach apply to metamemory
paradigms with dichotomous test performance (e.g., recall or no recall) and may have limited validity for typical metacomprehension paradigms. The computation of $D_a$ is a function of $X$ and $Y$ coordinates, where $X$ is the probability of one outcome (e.g., successful recall) for a given judgment level and $Y$ is the probability of the other outcome (e.g., failed recall) at that judgment level. Each item used to compute those probabilities has its own independently rated judgment, unlike metacomprehension paradigms where a judgment applies only to the continuous aggregate score on multi-item tests. It is unclear how this $D_a$ approach could be validly applied to continuous performance scales. Recoding continuous scores into dichotomous categories (e.g., by computing a median split) would discard much of the magnitude information just as Gamma does. In addition, Benjamin and Diaz (2008) show failed linearity tests for Gamma and Pearson at their minimum and maximum boundaries of $+1.0$ and $-1.0$ when using dichotomous performance measures. However, when performance is measured on a continuous interval scale with a range of values, extreme Pearson values are uncommon, because they require predicting the precise direction and magnitude of each relative difference. Finally, the recommended design for computing $D_a$ entails twenty observations per judgment level per participant. As noted above, this is untenable within the prevailing metacomprehension paradigm where readers are being asked to reflect on their comprehension of the complex relations among ideas from each text as a whole.

The magnitude information ignored by Gamma (and by $D_a$ measures used with nondichotomous performance measures) will tend to reduce the number of observed unique accuracy scores, increase standard deviations, and lead to ceiling/floor effects due to more observations being pushed to the maximum and minimum values of $\pm1.0$. For example, consider the results of one metacomprehension study (Wiley et al., 2017: Experiment 1) for which both judgments and performance were assessed on 6-point scales. In a sample of eighty-three participants, there were only twenty unique values for relative metacomprehension accuracy as computed with Gamma, including sixteen scores of 0, twenty-two scores of $+1.0$ and thirteen scores of $-1.0$. In contrast, there were sixty-eight unique values for relative metacomprehension accuracy as computed with Pearson, with only eight scores of 0, one score of $+1.0$ and zero scores of $-1.0$. Thus, Pearson provided a more continuous, normally distributed, and sensitive metric of covariance between predictions and test scores. The mean Gamma and Pearson were similar at 0.16 and 0.14 but the medians (0.00 vs. 0.13) and modes (1.00 vs. 0.00) reflect abnormal distribution for Gamma scores that are problematic for use as an outcome measure in tests of means. Gamma also resulted in larger standard deviations (0.15), as did $D_a$ (0.41) than did Pearson (0.11), producing smaller effect sizes and less theoretically coherent results across the three experiments for both Gamma and $D_a$ than for Pearson.

The prior observation that Gamma ignores potentially meaningful differences in judgment magnitude (Schwartz & Metcalfe, 1994) becomes doubly problematic when meaningful differences in test performance magnitude are also ignored. Imagine two different people both score a 0 percent, 1 percent, 99 percent, and 100 percent on four different tests. Thus, they had two tests they did almost as poorly as possible on and two they did almost as well as possible on. Of the six possible pairs
of tests, four pairs have huge differences that likely reflect reliable and meaningful differences in level of understanding, while Test 1 versus Test 2, and Test 3 versus Test 4 show the smallest possible 1 percent difference that is probably not meaningful or reliable. Now imagine that Person 1 gave the judgment predictions of 1 percent, 0 percent, 100 percent, and 99 percent. They were almost perfectly correct in predicting the four biggest pairwise differences that reflect real differences in understanding, and only failed to predict the relative order in the two pairs with the tiny 1 percent difference in understanding. In contrast, imagine Person 2 gave predictions of 0 percent, 99 percent, 1 percent, and 100 percent. Although Person 2 predicted the two tiny differences in performance, they were only at random chance (two of four) in predicting which of the two they would almost ace and which they would get near the lowest score possible. Arguably, any valid measure of relative accuracy should give Person 1 a much higher accuracy score than Person 2, which Pearson does (0.99 versus 0.02), but Gamma treats them as identically accurate (both 0.33) because both people had four concordances and two discordances. This makes Gamma more influenced than Pearson by the smallest least reliable differences in test performance. Once again, this problem for Gamma is particular to the complexities of text metacomprehension research because dichotomous memory tests have only one possible magnitude of difference in performance.

A conservative approach followed by increasing numbers of metacomprehension researchers is to compute and report multiple measures of relative accuracy including Pearson and Gamma (and $D_a$ when possible and appropriate). Often these alternative measures show similar patterns and reporting each allows the reader to make connections to the existing literature. However, when they diverge, it should not be presumed that the results are unreliable, but rather it could be that the methodological paradigm makes some measures more reliable and valid than others. With continuous performance measures and judgments made about that aggregate performance, Pearson may be the more psychologically valid and statistically powerful approach. Our own work in metacomprehension generally finds that Pearson scores provide more observations, more unique values, fewer extreme scores, lower standard deviations, and more normal distributions that are appropriate for testing differences in accuracy between conditions in general linear models.

**Cues That Predict Comprehension of Text**

In addition to using a measure of monitoring accuracy that is sensitive to readers’ online experiences, monitoring will theoretically be more effective when the cues that readers use as a basis for their judgments are more diagnostic for predicting comprehension (Dunlosky, Mueller, & Thiede, 2016). As already noted, cues based in heuristics are unlikely to reflect differences in comprehension across a set of texts. In addition, the use of different types of cues that are tied to meta-experiences may not all be equal. A central premise of the situation-model approach to metacomprehension (Wiley, Griffin, & Thiede, 2016) is that some meta-experiences (i.e., cues) may be more indicative of one’s ability to recall or recognize explicitly stated text
information, and other meta-experiences may be more reflective of one’s ability to recognize valid inferences and connections merely implied by the text, to connect new information to prior knowledge, or to apply it to new situations (Rawson, Dunlosky, & Thiede, 2000; Wiley, Griffin, & Thiede, 2005). This notion is rooted in the prevailing theory about different levels of representation that are constructed during text comprehension (Kintsch, 1998). Kintsch’s framework distinguishes several levels at which text information is represented: the surface level where exact word-forms and syntax are represented; the text-base level where explicitly stated individual idea units are represented; and the situation-model level where the meaning of concepts and their relations with each other and with existing knowledge are represented as an integrated and coherent whole. The quality of one level of representation does not necessarily correspond to the others. Complete surface-level memory for words can be achieved without any comprehension of the relations among concepts, and surface-level information can quickly fade from memory, while a more conceptual understanding based in the situation-model level remains (Kintsch et al., 1990). Further, it is the situation-model level representation that is most important for comprehension (Kintsch, 1994; Wiley et al., 2005).

Based on these premises, the situation-model approach suggests that the most predictive cues for judging comprehension will be those that reflect the quality of the situation model that the reader has constructed. Such cues could include fluency and ease of processing when trying to build bridging inferences or to engage in end-of-paragraph wrap-up, or the experience of success or difficulty when trying to summarize or explain ideas from the text. Thiede and colleagues (2010) asked readers to report what cues they had used to guide their judgments of comprehension. Reported cues were classified into three broad types: superficial, memory-based, and comprehension-based. Superficial cues could include some of the general heuristic cues described above, such as familiarity with or interest in the topic of the text. These cues are not necessarily tied to any representation of the actual text being read. Memory-based cues entail being able to recall or restate parts of the text whereas comprehension-based cues entail being able to explain the meaning of a text. Some examples of the types of cues readers reported were being able to remember ideas in a text (memory-based), whether they thought they could explain the meaning of the text to someone else (comprehension-based), and whether they felt that the topic was boring (superficial). Thiede and colleagues (2010) found that comprehension-based cues, though rarely reported, were associated with the highest levels of metacomprehension accuracy. However, even memory-based cues led to accurate predictions of test performance in a delayed generation condition. Jaeger and Wiley (2014) also examined readers’ self-reported use of either comprehension-based or noncomprehension-based cues and found that use of comprehension-based cues led to higher metacomprehension accuracy. Further, Wiley, Jaeger and colleagues (2017) found that prompting readers to use comprehension-based cues (e.g., “Do you think you could explain the causal process of digestion to a friend?”) led to higher relative accuracy than prompting them to use more superficial and memory-based cues (e.g., “How much of the text do you feel like you would be able to recall?”). Getting readers to generate and attend to diagnostic cues is the crux of several instructional contexts that have showed improvements over typical levels of relative metacomprehension accuracy.
Effective Interventions for Improving Metacomprehension

There is now a substantial body of work that has demonstrated the effectiveness of several types of activities in improving relative metacomprehension accuracy. Four main approaches are discussed below. Most of the approaches, such as those shown in the shaded box in Figure 24.1, impose alternative or additional processing tasks designed to help students generate more diagnostic cues, whereas other approaches guide students toward selecting the most diagnostic cues from those that are available (Griffin et al., 2013; Wiley, Griffin et al., 2016; Wiley et al., 2016).

Each of these approaches can be thought of as supporting the use of diagnostic cues during monitoring in slightly different ways and their two paths (A and B) are illustrated in Figure 24.2. However, they are all generally consistent with the theoretical assumptions proposed by the situation-model approach. Although the majority of these studies have been done in laboratory contexts with undergraduate participants, several have now been done in classroom contexts and with younger students.

Delayed generation tasks. Activities using delayed generation tasks, including delayed keyword and delayed summarization tasks, have been shown to improve metacomprehension accuracy, presumably because they help readers to generate cues related to the situation model. In the original study using a delayed-generation paradigm, Thiede and Anderson (2003) found that relative metacomprehension accuracy was higher for students who wrote summaries after a delay and then judged their performance than for groups who either wrote a summary immediately after reading or who wrote no summaries. Similarly, Thiede, Anderson, and Therriault (2003) showed that a delayed keyword generation task, where students were asked to list keywords that captured the gist of a text, produced the same boost in relative metacomprehension accuracy relative to generating keywords immediately after reading or not generating keywords. In addition to observing benefits for relative metacomprehension accuracy, Thiede and colleagues (2003) also found that delayed keyword generation led to more effective self-regulation of study. Participants in the delayed keyword generation condition were more likely to select texts for restudy based on perceptions of how well the texts had been learned during initial reading than participants in the immediate and no generation conditions. Although the three groups did not differ in performance on the initial comprehension test, the delayed keyword generation group outperformed both other groups on the final comprehension test, which occurred after restudy, and made the greatest gains in performance between the first and final comprehension tests. These findings suggest that delayed keyword generation may not only improve relative metacomprehension accuracy but also lead to more effective restudy behavior and, ultimately, gains in learning.

Why do delayed generation tasks improve relative metacomprehension accuracy? Both components, the generation activity and the delay, play a role. First, generating a summary or keywords may allow a reader to reflect on how successfully he or she is able to retrieve information during generation (see the modified feedback hypothesis described by Glenberg et al., 1987). Accordingly, a text may receive a high rating of comprehension if the person is able to retrieve a great deal of information about the
text during generation; whereas, a text may receive a low rating of comprehension if the person struggles to retrieve information about the text. Second, the timing of the generation task is critical. Kintsch and colleagues (1990) showed that surface memory for text decays over time, whereas the situation model is robust to such decay. When writing a summary immediately after reading, a person may have easy access to their surface model (or episodic memory for the text) and can use this information to generate a summary. However, relying on the surface model as a basis for a judgment of comprehension fails to improve accuracy because performance on the immediate summary task and the later comprehension test are determined by different levels of representation. In contrast, when writing a summary after a delay, the findings by Kintsch and colleagues (1990) suggest that a person will likely have relatively greater access to the situation model of a text. Thus, using the experience of writing a summary after a delay as a basis of a judgment of comprehension improves accuracy because performance in both the delayed summary task and on the comprehension test are both based in the situation model.

Thiede, Dunlosky, and colleagues (2005) tested this explanation for delayed generation effects by independently varying several different features of generation tasks, including the types of delay and the nature of the generation task. Consistent with the situation-model approach, the critical feature for improving metacomprehension accuracy was a delay between reading and generation, because only this delay forces readers to rely on their situation model to generate the keywords and thus to generate cues related to the quality of that situation model. Other delays (between generation and judgment or between the generation task for one text and the next) did not improve accuracy. Performing nongenerative tasks at a delay, such as reading a list of provided keyword or prompting readers to “think about the text,” also did not improve accuracy. The key to producing better monitoring was making readers perform a specific type of generative self-test.

Improvements in relative monitoring accuracy due to engaging in delayed generation tasks have now been replicated by several labs in a variety of contexts. Anderson and Thiede (2008) found benefits of delayed summary generation in a sample of US college students. Shiu and Chen (2013) replicated delayed keyword effects with a sample of Chinese college students. Thiede and colleagues (2010) showed benefits of delayed summary generation for both typical and at-risk college readers, and found that the delay changed the nature of the cues produced during keyword generation such that they became far more predictive of performance on inference tests. In addition, benefits of delayed keyword tasks have now been shown with younger students reading expository texts in the context of science lessons. Several studies have shown that even middle school students can benefit from delayed keyword generation tasks when making judgments of comprehension (de Bruin et al., 2011; Thiede et al., 2012; Thiede et al., 2017). However, de Bruin et al. (2011) found that 4th graders were not able to benefit from a delayed keyword activity, which suggests that delayed generation benefits may not extend to younger readers.

The fact that a delay by itself is insufficient to improve judgment accuracy underscores a critical difference between metacomprehension of text and metamemory for simpler materials, like paired associates. In metamemory paradigms, a delay
before judgment does notably improve accuracy (Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011). The judgment prompt for paired associates is essentially the same as the testing prompt. Thus, the learner can attempt to self-test and generate a test answer as part of the judgment process, which produces diagnostic cues for predicting that performance. However, in metacomprehension paradigms the judgment is about comprehension of complex, interrelated concepts. Unlike in metamemory paradigms, there is no one-to-one mapping between the judgment and testing prompts because the judgment prompts are far broader and more abstract than the actual test items. Thus, a delay between reading and judgment is not sufficient to improve relative metacomprehension accuracy. Rather, the delay merely moderates whether the generation task produces cues that reflect the situation model rather than surface memory.

**Rereading.** Another activity that can help students to improve their metacomprehension accuracy is rereading a text before making judgments of comprehension. Sizable increases in monitoring accuracy have been observed as a result of prompting college students to read each text in a set twice in succession before making their judgments (Dunlosky & Rawson, 2005; Rawson et al. 2000). However, rereading was not found to be an effective activity for improving metacomprehension accuracy among 7th graders presented with a set of expository texts as part of their science lessons (Redford et al., 2012). Why might rereading be an effective strategy? Dunlosky and Rawson (2005) have noted that “rereading may afford more resources for comprehension monitoring” (p. 51). During an immediate second reading, many of the subprocesses involved in reading do not need to be reexecuted (e.g., Millis, Simon, & tenBroeck, 1998; Perfetti, 1985). Importantly, Dunlosky and Rawson (2005) found greater benefits from immediate rereading than from delayed rereading.

This reduction in the need for low-level text processing during immediate rereading allows readers to focus more attention at the meta-level during a second pass and may have implications for readers who struggle to monitor during reading. To test this hypothesis, Griffin and colleagues (2008) examined effects of rereading as a function of individual differences in both comprehension skill and working memory capacity (WMC). When readers were low in comprehension skill or WMC, their metacomprehension accuracy after a single reading was limited. However, following immediate rereading their metacomprehension improved such that they were just as accurate as readers with more comprehension skill or WMC. Griffin and colleagues (2008) argued that rereading helps by allowing readers a second chance to attend to monitoring cues, which can be especially beneficial for readers whose limited comprehension skill or attentional resources prevent them from attending to anything but low-level text processing during the first reading. Unlike the benefits of delayed generation discussed above, the benefits of rereading are not rooted in drawing the reader’s attention to situation model-based cues per se. Instead, by decreasing reading-related processing demands, rereading may be beneficial because it helps readers (particularly struggling ones) to be able to attend to
any meta-level experiences at all rather than rely on heuristic cues that are just as available before and after reading.

**Explanation and concept map generation tasks.** While delayed generation activities may prompt attention to an already constructed situation-model, self-explanation may encourage efforts to construct a situation model during reading and thereby help readers to generate relevant cues about that level of comprehension. Self-explanation is the activity of explaining to oneself during reading in an effort to understand and integrate new information (Chi, 2000). Griffin and colleagues (2008) tested the hypothesis that self-explanation during rereading might improve relative accuracy by generating meta-experiences such as a subjective sense of difficulty in generating an explanation or a sense of explanatory coherence. Participants were randomly assigned to one of three groups. One group read each text once, a second group read each text twice, and a third group read each text twice and engaged in self-explanation of connections between parts of the text during rereading. Metacomprehension accuracy was higher in the rereading condition relative to the single reading condition, but the group that self-explained during rereading had significantly better metacomprehension accuracy than the other two groups. Fukaya (2013) has also replicated the benefits of having participants generate explanations using a set of illustrated texts about “How Things Work.” Across two studies, engaging in generation of explanations following reading led to higher accuracy than being told to “expect” to have to generate explanations or generating keywords immediately.

Griffin and colleagues (2008) found that even readers with limited resources, due to low WMC or low comprehension skill, benefited from self-explanation above mere rereading. They suggested that the benefits from self-explanation on monitoring accuracy may be due to specific types of meta-experiences and cues generated by the task. However, in this study, participants always engaged in self-explanation during a second reading. Further, the rereading condition suggested that a second reading was critical for these struggling readers. So, it is plausible that struggling readers will only benefit from the types of cues generated by tasks like self-explanation if they are able to attend to meta-experiences in general, which is made possible by a second reading.

More recent work has begun to explore the use of concept-mapping tasks as learning activities and artifacts that might help to improve the metacomprehension accuracy of readers with resource limitations. Concept-mapping was considered to be a promising candidate for a learning activity to the extent that it may help readers deal with the competing demands of reading and monitoring. Since the text is available during the mapping activity, and the activity creates a visual representation of the situation model, concept-mapping tasks may be especially appropriate for students who have limited processing resources (Stensvold & Wilson, 1990). Indeed, in a review of the literature, Nesbit and Adesope (2006) concluded that concept-mapping tasks were the most effective activities for improving learning from text for younger, less-skilled, or at-risk readers.
In addition, constructing concept maps should help readers generate diagnostic comprehension cues. Weinstein and Mayer (1986) suggested that instructing students to create concept maps of texts during reading helps them to identify the connections among concepts, which improves comprehension. From this perspective, concept-mapping is similar to self-explanation, as both tasks should help readers construct and attend to the underlying causal models of the subject matter, thereby generating meta-experiences that reflect the quality of their mental representation of those causal models. Consistent with this reasoning, Thiede and colleagues (2010: Experiment 2) tested the effectiveness of a set of concept-mapping activities on improving metacomprehension accuracy in a sample of college students enrolled in a remedial reading course. In a within-subjects design, initial relative metacomprehension accuracy was obtained using the standard paradigm. That is, students read a series of texts, judged their comprehension of each text, and then completed a comprehension test for each text. Participants then received lessons on how to construct causal concept maps from short scientific texts and completed the standard metacomprehension paradigm again, but this time they constructed concept maps while reading. Constructing concept maps while reading improved both comprehension and relative metacomprehension accuracy.

Similar studies with middle school students have shown that simply prompting students to generate concept maps, or to fill in concept maps immediately after reading, does not lead to higher levels of metacomprehension accuracy among younger readers (Redford et al., 2012; van Loon et al., 2014). However greater improvements were found when younger readers were shown a good example and given instruction in how to create and use concept maps for understanding (Redford et al., 2012) and when readers were given a template to fill in after a delay (van Loon et al., 2014). Under these conditions, concept-map generation activities seemed to help young readers use cues based in their situation model when making judgments of comprehension.

The benefits of concept-map generation are consistent with benefits found with other tasks that may also direct the reader’s attention to the situation model. Thomas and McDaniel (2007) found that sentence sorting (a task which relies on the reader’s situation model) improves accuracy for predicting performance on thematic questions, while letter insertion (filling in missing letters in words, which does not rely on the reader’s situation model) improves accuracy for predicting performance on detail questions. These different effects for comprehension and detail questions harken back to the difference between metacomprehension and metamemory for text paradigms discussed above. Tasks that generate meta-experiences tied to the situation model should lead to benefits specifically when readers are asked to predict performance on comprehension questions rather than detail (memory) questions because only the former are likely to depend directly on the quality of the situation model. In contrast, answering detail questions is more likely to depend on surface memory and text-base representations. When tests consist of detailed memory questions, then monitoring should benefit from tasks that generate meta-experiences related to those levels of representation.
Learning goals and cue selection. A common feature of the manipulations discussed so far is that they directly prompt the reader to engage in additional processing tasks in order to generate the meta-experiences that readers can draw on as diagnostic cues for the judgments of comprehension. More recently, studies have been exploring instructional conditions that may help readers select the most diagnostic cues for comprehension monitoring. One example is giving students the expectancy that their comprehension will be assessed with inference tests rather than memory tests. This work takes as its starting point the premise that students need to have a valid reference for what it means to “comprehend” an expository text in order to make judgments that predict comprehension (Wiley et al., 2005). Without specific instructions about what comprehension entails, what goals for reading should be, and what comprehension tests will be like, students may select memory-based cues instead of comprehension-based cues as the basis for their judgments.

Encoding the exact ideas from a text into memory may be a student’s default setting for reading. Many studies have demonstrated that people tend to be better at predicting their ability to answer detail questions, or questions that rely on their memory for specific terms, as opposed to comprehension questions that rely on the ability to integrate, connect, or apply information from the text (Dunlosky, Rawson, & Middleton, 2005; Ozuru, Kurby, & McNamara, 2012; Rawson & Dunlosky, 2007; Thiede et al., 2012; Thomas & McDaniel, 2007; Weaver & Bryant, 1995; Wiley, Griffin, & Thiede, 2008).

Using memory-based cues as a basis for JOLs from a text will generally be an effective strategy if future tests ask for specific terms, details, or ideas directly stated by a text. Further, this behavior is certainly important for some subject matters and learning contexts. However, if tests require students to gain conceptual understanding, for example of scientific processes and phenomena from expository text, then it is important to prompt students to override the “reading for memory” setting. In order to engage in accurate monitoring, readers need to appreciate that their goal for reading is understanding how or why a phenomenon or process occurs, and that they will need to make connections and causal inferences across sentences. Instilling appropriate reading goals may be particularly important when readers are tasked with understanding science texts.

Thiede, Wiley, and Griffin (2011) tested the hypothesis that test expectancy impacts metacomprehension accuracy by manipulating whether students were provided with an explicit statement about the nature of comprehension, an explicit statement about the nature of the final test items they should expect, and example test items for a practice text (on a different topic than the target texts). This test-expectancy manipulation was highly effective for a sample of graduate students in educational psychology. Griffin and colleagues (2008) found similar improvements among a sample of undergraduates. Participants in the comprehension expectancy condition were better able to predict performance on comprehension items for target texts over no-expectancy and memory-expectancy conditions. Further, when a self-explanation prompt was combined with comprehension test expectancy, the benefits were additive and led to the highest levels of comprehension monitoring accuracy. This combined benefit could reflect the two distinct paths of monitoring strategies
shown in Figure 24.2. An added cognitive activity like self-explanation allows learners to generate useful meta-experiences (Path A), while the comprehension test-expectancy allows learners to select the more diagnostic experiences available to them when making their judgments (Path B).

It is encouraging that similar benefits have also been found for this combined manipulation in a real classroom context (Wiley et al., 2016). Students in a research methods course were randomly assigned to receive the same combined test-expectancy and self-explanation intervention used in Wiley and colleagues (2008). Both the combined-manipulation group and a control group read six texts on research methods topics and made monitoring judgments. A week later, the groups were reminded of their judgments and given time to restudy before taking quizzes on all the texts. Not only did students who received the combined test-expectancy and self-explanation instruction have higher monitoring accuracy but they were also more likely to restudy the texts in a strategic manner (rather than just reread in order) and their restudy was more effective in producing learning gains evidenced by their quiz scores.

Comprehension-based curricular experience. The test-expectancy studies have attempted to give students a better understanding of appropriate goals for comprehending expository science texts as part of experimental manipulations. However, if prior literacy instruction already provides students with this knowledge, then they should be in a better position to engage in accurate comprehension monitoring. Indeed, Thiede and colleagues (2012) found that 7th and 8th grade students whose early literacy education focused on deep understanding and experience with inference tests demonstrated better metacomprehension accuracy than students with more typical schooling experience. Moreover, superior monitoring accuracy led to better decisions about which texts to restudy and produced significantly better overall comprehension. A potential explanation for this result is that prior experiences with inference tests created an expectancy that led students to select more diagnostic cues when judging their comprehension. Commander and colleagues (2014) reported a similar advantage in metacomprehension accuracy for Chinese college students over American college students. They suggested this advantage could be due to the emphasis that Chinese reading instruction puts on “meaning making.”

Taken together, the above studies suggest that when readers rely on appropriate cues for judging their comprehension, either by engaging in activities that help them to generate more diagnostic cues or by having reading goals or curricular experience that direct them toward the selection of more diagnostic cues, they are able to engage in more accurate comprehension monitoring.

Conclusions and Future Directions

Without special instructions or activities, most students tend to have very poor metacomprehension accuracy, and several reviews have reported average levels of metacomprehension accuracy around .27 in baseline conditions
(Dunlosky & Lipko, 2007; Maki, 1998a, 1998b; Thiede et al., 2009). However, a substantial body of evidence emerging from studies inspired by the situation-model approach to metacomprehension indicates that there are a number of conditions that can lead to significant improvements in metacomprehension accuracy. The situation-model approach suggests that the most predictive cues for judging metacomprehension will be those that reflect the quality of the situation model that the reader has constructed. Such cues could include a sense of fluency and ease of processing when trying to build bridging inferences or engage in end-of-paragraph wrap-up, or the experience of success or difficulty when trying to summarize or explain ideas from the text. A growing number of studies have demonstrated that conditions that prompt readers to rely on appropriate cues for judging their comprehension, either by engaging in activities that help them to generate meta-experiences that can serve as diagnostic cues or by having reading goals or curricular experience that direct them toward the selection of more diagnostic cues, can successfully improve students’ comprehension monitoring.

The results of several studies using instructional and test-expectancy manipulations suggest that it is important for teachers to give students clear goals for reading and clear expectations about the nature of test items. There seems to be sufficient support to recommend instruction aimed at teaching students about the difference between what it means to remember a text versus what it means to comprehend the ideas and arguments a text is making. In addition, students could be encouraged to keep this distinction in mind while reading, using it to help them make better judgments of whether they have actually comprehended a text. This ability to judge comprehension should then positively impact learners’ decisions about what to restudy. The results of several studies also suggest that students could be prompted to engage in tasks like concept-mapping and self-explanation, not merely to directly improve comprehension but also to improve metacomprehension. Of the various effective activity manipulations, delayed keywords and summaries appear to be the simplest to implement in the classroom. However, such simple tasks may lack the direct benefits to comprehension itself that more involved tasks, such as self-explanation and concept mapping, have been shown to produce.

Although a substantial number of studies have been able to show improvements in metacomprehension accuracy, there are still several important directions for future research.

First, it is important to note that most of this work has examined comprehension monitoring from plain expository texts. However, in many contexts expository texts are accompanied by instructional adjuncts such as images, animations, and analogies. These adjuncts are often included with the goal of supporting better student understanding. While these adjuncts could potentially improve both comprehension and metacomprehension outcomes, they have also been suggested to lead to illusions of understanding (Jaeger & Wiley, 2014, 2015; Wiley, Jaeger, et al., 2017; Wiley, Sarmento, et al., 2017). More work is needed to understand how teachers can help
students gain only the benefits of including instructional features such as visualizations and analogical examples.

Further attention to the fact that almost none of the metacomprehension accuracy studies has explicitly informed readers about the potential metacognitive purpose of the various additional tasks or instructions is also needed. It is likely that readers would assume the focus of additional tasks is on comprehension itself and readers may not treat the tasks as part of a metacognitive strategy. There is no obvious reason that overt awareness of this metacognitive utility should undermine the observed benefits. It may even boost the effects, but this is an untested empirical question. In addition, although several studies have now been done in classroom contexts and with younger students, much more work is needed to test to what extent findings obtained with undergraduate samples and in laboratory contexts will generalize to other samples and contexts.

The focus of this chapter has been on relative metacomprehension accuracy, but more work is needed that measures both absolute and relative accuracy because they are likely to have distinct implications. However, it is crucial that the findings be evaluated independently for each measure. Researchers should avoid lumping the two sets of results together as though they are alternative convergent measures of the same phenomena. Divergent results from the two measures are to be expected, and when different measures fail to show the same result, this should not be seen as a failure to replicate. Instead, more careful attention is needed to how each might contribute to effective SRL, and what conditions are likely to cause improvements in each of the measures.

Finally, there is a great need for more research on how and when improvements in metacomprehension accuracy lead to improvements in self-regulated study and learning outcomes. More work is needed to understand the relation between improving metacomprehension accuracy and causing positive changes in study behaviors that improve learning outcomes. A major obstacle to achieving these goals is that it appears impossible for the same study to assess the full causal chain from predictive monitoring accuracy, to restudy, and to final level of comprehension. If readers make judgments prior to restudy, then the final test performance cannot serve as a valid measure of metacomprehension. Effective restudy should mean that learners are selectively restudying each text relative to how much their learning would benefit. Thus, effective restudy is likely to alter the relative comprehension levels among the texts and thus reduce how well they correlate with the pre-restudy judgments. Readers who most effectively use their metacomprehension judgments to guide restudy will appear to have lower metacomprehension accuracy if computed using post-restudy tests. A seeming solution to this problem is to test for comprehension prior to restudy regulation and use this performance to assess metacomprehension accuracy. However, that testing gives learners test experience and implicit feedback about performance on objective tests that they can use to guide their restudy. That undermines the ability to draw any inferences about whether actual metacognitive monitoring during initial learning is playing a critical role in later study regulation. Thus, researchers will need (and hopefully reviewers will appreciate this need) to piece
together the chain of causality from initial experiments that show the impact of a manipulation on monitoring accuracy and follow-up experiments that show the same manipulation impacting restudy regulation and ultimately learning. Similarly, much more work is needed to understand how to achieve long-term benefits from interventions. These kinds of projects are important for providing empirical support in order to make the strongest recommendations for education.

References


