This article should be cited as:
Griffin, T. D., Wiley, J., & Thiede, K. W. (2019). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(6)*, 1066-1092. http://dx.doi.org/10.1037/xlm0000634

The Effects of Comprehension-Test Expectancies on Metacomprehension Accuracy

Thomas D. Griffin & Jennifer Wiley

University of Illinois at Chicago

Keith W. Thiede

Boise State University

Author Note

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305B07460 and R305A160008. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

Correspondence should be sent to Thomas D. Griffin, Department of Psychology, University of Illinois at Chicago, 1007 W Harrison Street (M/C 285), Chicago IL, 60607; E-mail: tgriffin@uic.edu

Abstract

A set of four experiments assessed the effects of establishing a comprehension-test expectancy (in contrast to a memory-test expectancy) on relative metacomprehension accuracy. Typically readers show poor relative metacomprehension accuracy while learning from text (i.e., they are unable to discriminate topics they have understood well from topics they have understood poorly). In the first experiment, both readers who were given no test expectancy and those who were given a memory-test expectancy made judgments that were more predictive of performance on memory tests than inference tests. However, readers who were given a comprehension-test expectancy made judgments that were more predictive of inference-test performance. This effect was replicated and extended in two additional experiments that showed an effect of comprehension-test expectancy even when no example test items were provided, and when the expectancy was established only after reading. A fourth experiment showed that establishing a comprehension-test expectancy still had an effect on accuracy even when metacomprehension accuracy was already being improved via a self-explanation activity. The results show robust and reliable benefits to metacomprehension accuracy from a comprehension-test expectancy that serves as portable knowledge that learners can apply to monitoring future learning from text.

Keywords: metacomprehension, metacognition, test expectancy, monitoring accuracy, judgments of learning

The Effects of Comprehension-Test Expectancies on Metacomprehension Accuracy

The present research is concerned with the effect of test expectancies on monitoring accuracy while learning from text. Monitoring accuracy is defined as the ability to accurately predict how well one will do on a later test of studied material. A prototypical finding is that most readers lack the ability to accurately monitor their own comprehension, and in particular, that they are unable to discriminate topics they understand well from topics they understand less well (Dunlosky & Lipko, 2007; Maki, 1998). *Relative metacomprehension accuracy* serves as the measure that represents this discrimination ability; that is, how well the variance in a learner's judgments for an array of texts covaries with the variance in that learner's performance on tests for those texts. It is computed using intra-individual correlations between predictive judgments of comprehension and actual test performance for a set of texts. Several reviews of empirical work using this measure have demonstrated that average values for relative metacomprehension accuracy are generally only around .27 (whereas perfect accuracy would result in a positive 1.0 value; Griffin, Mielicki, & Wiley, in press; Maki, 1998; Thiede, Griffin, Wiley, & Redford, 2009). This is problematic as the ability to discriminate one's understanding among different topics is a critical skill for the effective self-regulation of learning and study behaviors. Poor relative metacomprehension accuracy leads readers to make sub-optimal choices such as failing to re-study poorly understood information while attempting to learn from texts (Maki, 1998; Thiede, Anderson, & Therriault, 2003; Wiley, Griffin, Jaeger, Jarosz, Cushen, & Thiede, 2016).

The central premise that is tested in this series of studies is whether providing a comprehension-test expectancy will impact relative metacomprehension accuracy. A comprehension-test expectancy means informing readers that future tests will assess their understanding and ability to make connections across ideas within a text, rather than simply the ability to remember ideas from a text. In the remainder of this introduction, the motivation for this work is explicated by considering theories of metacognitive monitoring that conceive of monitoring as a cue-based judgment process and articulating the importance of using diagnostic cues as part of such a judgment process. Text-processing theories are considered in order to identify which cues are most likely to be diagnostic specifically for comprehension outcomes, and prior empirical work is considered with respect to which approaches have been most effective at improving relative metacomprehension accuracy. The main prediction that follows from this theoretical and empirical overview is that a comprehension-test expectancy should improve relative metacomprehension accuracy.

Theories of Metacognitive Monitoring

Why are readers so poor at monitoring their own level of comprehension? In general, theories of metacognitive monitoring are inference-based approaches that characterize monitoring as a judgment process (Dunlosky, Mueller, & Thiede, 2013; Koriat, 1993; Schwartz, Benjamin, & Bjork, 1997). Inference-based approaches assume that people make monitoring judgments by inferring how potential cues are predictive of their performance. Cues can include features, properties or characteristics of the to-be-learned stimuli; characteristics of the learner; perceptions of the learning context; or subjective experiences triggered by reading processes and learning episodes. According to inference-based approaches, poor monitoring accuracy results from the use of inappropriate cues as the basis for monitoring judgments. The cue-utilization framework (Koriat, 1997) suggests that readers will have poor monitoring accuracy when they base their judgments on cues that are not valid predictors of the performance being measured. Similarly, applying the ideas of Brunswik (1956), monitoring accuracy will be poor when the cues that are used as a basis for judgments are not diagnostic of actual comprehension. In the context of these theories, one can differentiate potential cues as being more or less diagnostic and appropriate to use as a basis for comprehension judgments.

Flavell's (1979) original conception of metacognitive monitoring proposed that subjective *meta-experiences* serve as the basis for accurate judgments, and that such experiences are generated during encoding or use of knowledge. Based on Flavell's model, Griffin, Wiley and Salas (2013) proposed a distinction between more appropriate (or more diagnostic) judgments, based on Flavell's meta-experience cues, and less appropriate (or less diagnostic) judgments of comprehension, based on heuristic cues. Heuristic-cue-based judgments are less appropriate and less diagnostic because they do not actually reflect "monitoring" during specific learning episodes, but rather reflect non-experiential presumptions. Some examples of common heuristic cues are generalized self-efficacy beliefs; perceptions of ability, topic interest and familiarity; or perceptions based in features of the stimuli, such as text length or font size, rather than cues that more directly reflect the learning experiences during study of each text. Readers who self-report using heuristic cues have been shown to have poorer relative metacomprehension accuracy (Thiede, Griffin, Wiley, & Anderson, 2010). Similarly, when readers are unable to attend to experience-based cues, and are forced to default to heuristic cues due to working-memory limitations, their relative metacomprehension accuracy has been shown to suffer (Griffin, Wiley, & Thiede, 2008).

Further, theories of text comprehension suggest that using just any type of experience-based cue is not enough to ensure accurate metacomprehension (Rawson, Dunlosky, & Thiede, 2000; Weaver, Bryant, & Burns, 1995; Wiley, Griffin, & Thiede, 2005). Kintsch's theoretical framework (1998) encapsulates the general view that text processing entails representation at multiple levels, and some experience-based cues may reflect processing at one level but not others. The surface level involves a memory representation of the exact words that are read, while the textbase level encodes the meaning of individual propositions. At the situation-model level, important connective and causal inferences are represented via integration of multiple text propositions with prior knowledge, and via generation of implicit relations. Thus, it is this level that represents the readers' mental model of the situation being described by the text. When a student is reading expository science texts with the goal to develop a mental model of causal processes and systems, then only the situation-model level of representation will be diagnostic of comprehension (Mayer, 1989; Otero, Leon, & Graesser, 2002; Wiley & Myers, 2003). The situation model determines how well readers can perform on comprehension tests that require them to apply information from explanatory expository texts in novel contexts, and to generate and verify possible inferences that follow from the text (Kintsch, 1994; Mayer, 1989). Thus, text comprehension research has contrasted between performance on inference-based comprehension tests and performance on memory-based tests that entail recall or recognition of explicitly stated text information and only require use of a surface-level representation, or at most the text-base, if paraphrases and synonyms are used (e.g., McNamara, Kintsch, Songer, & Kintsch, 1996)

By integrating text comprehension and metacognitive monitoring frameworks, metacomprehension researchers have argued that readers will be unable to accurately predict performance on comprehension tests (as opposed to memory tests) unless they base their judgments on experiences generated during the creation or use of their situation-model-level representations (Rawson, Dunlosky, & Thiede, 2000; Weaver, Bryant, & Burns, 1995; Wiley, Griffin, & Thiede, 2005). Cues based in such experiences could include a sense of coherence during self-explanation, a sense of fluency when attempting to summarize after a delay, or a sense of confusion when encountering a conclusion that supposedly follows from prior text. This perspective can be characterized as a *situation-model-cues* approach to metacomprehension accuracy. It suggests that inaccurate judgments of comprehension can occur either when readers use heuristic cues as discussed above, or when readers use meta-experiences that are only tied to a lower-level representation of the text (e.g., the ability to remember specific verbatim details or immediately recall a text), rather than experiences tied to the situation model. In support of this prediction, Thiede et al. (2010) found that readers tend to default to heuristic, superficial, or memory-based cues rather than comprehension-based cues when judging their own understanding, and also that judgments based in these cues were less accurate than judgments based in situation-model cues for predicting performance on tests of comprehension. Jaeger and Wiley (2014) replicated this result, and showed that readers' self-reported use of situation-model-based cues better predicted relative performance on inference tests but not on memory tests. The fact that different cue types were predictive of performance on different test types demonstrates the need to distinguish between two types of relative monitoring accuracy, namely metamemory versus metacomprehension accuracy. These two constructs are operationalized as the intra-individual correlation between a set of judgments and performance on a set of either memory or comprehension tests, respectively.

Improving Metacomprehension Accuracy by Manipulating Cognitive Processes

Most evidence supporting the situation-model-cues approach comes from studies that have directly manipulated readers' cognitive processing of the to-be-learned information. This has been done by requiring readers to engage in additional encoding or generative tasks designed to impact the construction, use, and access of situation-model-level text representations. These instructional tasks have directly manipulated processing of the to-be-learned material either during reading or at the time of judgment via supplemental tasks, such as delayed keyword generation or delayed summary generation (Thiede & Anderson, 2003; Thiede, Anderson, & Therriault, 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005), explanation during reading (Griffin et al., 2008), drawing or concept mapping during reading (Fukaya, 2013; Redford, Thiede, Wiley, & Griffin, 2012; Thiede et al., 2010; Van Loon et al., 2014), re-reading (Dunlosky & Rawson 2005; Griffin et al., 2008; Rawson et al., 2000), and text unscrambling (versus letter insertion, Thomas & McDaniel, 2007). Each of these instructional activities has been shown to improve relative metacomprehension accuracy above baseline levels. Interpreted in terms of models of metacognitive monitoring (e.g., Flavell, 1979; Koriat, 1997), these activities lead to more accurate judgments because they require additional cognitive processing that generates meta-experiences which can serve as situation-model-based cues. Explanation, drawing, and concept mapping entail constructing a situation-model of the phenomena, which makes situation-model-based cues more accessible at the time of judgment. Alternatively, delayed generation tasks improve monitoring because a delay after reading causes the surface information from each text to decay (Kintsch, Welsch, Schmalhofer, & Zimny, 1990), which forces the reader to rely upon their situation-model to perform the generation task. This produces more situation-model-based cues that are accessible at the time of judgment. Because these activities are not reader-initiated, but experimentally-required, the improvements to relative metacomprehension accuracy can be viewed as a byproduct of the instructional activities that produce diagnostic meta-experiences. The boost in accuracy can be produced without readers needing to strategically select which experiences or sources of information to use as a basis for their judgments. Rather, they benefit as a byproduct of the additional activities they are instructed to engage in.

For example, Thiede et al. (2010) showed that a delayed-summary activity increased relative metacomprehension accuracy without altering the type of cues that readers think they are using. From the readers' perspective, any experiences tied to generating summaries would be a similar "recall" cue type, regardless of whether the task was performed at a delay. Even though readers reported using similar "recall" cues following both immediate and delayed summaries, the recall cues became more predictive when the summary task was delayed. Because memory for a text loses surface detail over time while retaining the gist (Kintsch, Welsch, Schmalhofer, & Zimny, 1990), the delayed summaries provided cues that were better predictors of performance on comprehension tests. Thus, even when readers in different conditions approached the judgment task the same way, and reported relying on similar experiences, their judgments ended up being differentially predictive because of the specific features of the supplementary activities that were manipulated by the experimenter.

When accuracy is improved via experimental manipulations that alter the encoding of specific texts, or change the context under which judgments are made, there is no reason to think that learners have acquired any kind of transferable knowledge or skill that they could apply when trying to gauge their metacomprehension on a new set of texts. If readers are not being given any information that could lead them to modify their metacognitive approach or strategies on their own, then these activity manipulations are likely to only improve judgment accuracy for the specific texts where supplemental processing activities are required.

Improving Metacomprehension by Altering Metacognitive Goals

An alternative approach to directly manipulating cognitive processing of the text information is to make the learner more active in improving their relative metacomprehension accuracy by giving them a metacognitive goal that they can later apply when making judgments.

The present experiments test whether informing readers about the general nature of the upcoming tests can lead to improvements in relative metacomprehension accuracy. This approach could improve relative metacomprehension accuracy without explicitly instructing readers to engage in an activity that alters the encoding of the texts.

Typical students may view the concept of reading comprehension more in terms of memory for the text than understanding of text (Wiley et al., 2005), and may anticipate tests that only require recall or recognition. They may thus default to monitoring judgments based on memory-related cues (Thiede et al., 2010). If readers can be given an appropriate general expectation about the nature of upcoming tests as requiring the ability to make inferences and draw connections among ideas presented in a text, then readers might apply this knowledge toward utilizing judgment cues that will more accurately predict comprehension test performance with items that require such inferences. By not imposing any additional tasks beyond unstructured reading, any benefits would reflect the learners' application of an appropriate general expectation to better regulate their metacognitive processes on a new set of texts. A pedagogical benefit of altering learners' general metacognitive goals is that it may be more likely to be adopted in authentic learning contexts than prior manipulations that add to students' workload by requiring supplemental processing tasks during reading.

Two conceptual distinctions can be made between manipulations that improve monitoring accuracy via additional activities versus manipulations that attempt to establish a metacognitive goal. One distinction, as described in the prior section, is whether any changes to relevant processes are reader-initiated or experimentally-required. The second concerns whether the benefits are due to changes in text processing or changes in the judgment process. This latter distinction maps onto the distinction that Flavell (1979) made between *cognitive strategies* versus *metacognitive strategies*. Cognitive strategies refer to the actual operations that a learner engages in while processing the target information which automatically generate meta-experiences. Metacognitive strategies refer to meta-level goals and strategies that a learner needs to actively apply in order to either generate or make optimal use of meta-experiences for monitoring their learning. Prior manipulations have instructed readers to engage in additional activities, which directly alter readers' cognitive actions during text processing, and have the byproduct of improving the accuracy of monitoring judgments. These improvements in monitoring accuracy did not require that learners initiate any changes to their typical monitoring processes. Simply performing the instructed tasks was sufficient. In contrast, the current manipulation was designed to modify readers' general metacognitive goals (i.e., knowledge about the type of understanding they will need to achieve). In order for this to improve monitoring accuracy, readers would need to play a more active role by applying that general knowledge to modify their approach when predicting comprehension for new set of specific texts. In the present set of studies, we used a test-expectancy manipulation that either led readers to expect memory-based tests or inference-based tests in order to modify readers' general metacognitive goals.

Manipulating Test Expectancies

Test expectancies can be established in a number of ways, including by providing explicit descriptions about the type of tests that will be given or by giving readers experience with example tests. Most prior metacomprehension studies that have used practice tests have given initial tests on the same reading material that was assessed by the final tests. Although some studies have found increased relative metacomprehension accuracy following practice tests on the target material (e.g., Maki & Serra, 1992; Maki, 1998), these studies cannot isolate the effect of test *expectancies*. Learners get implicit feedback from their performance on same-text practice tests (Glenberg, Sanocki, Epstein, & Morris, 1987), and may base judgments on their performance for the prior practice trials (Dunlosky & Metcalfe, 2009). Thus, practice tests on the same topics essentially turns judgments from predictions into postdictions, where the reader can bypass monitoring of meta-experiences generated during the comprehension process, and can simply use past test performance to predict future performance. Prior work has shown that postdictions made after taking a test are generally more accurate than a priori predictions (Griffin et al., 2013; Pierce & Smith, 2001). Also, like other activity-based manipulations, practice tests can have a direct effect on encoding and processing of text information, as demonstrated in research on the well-established *testing effect* (e.g., Roediger & Karpicke, 2006). Thus, in the present studies, it was important to present example test items for *a different set of texts* than the ones on which relative metacomprehension accuracy would be assessed.

Another way to establish test expectancies is to inform students about the general nature of the tests they will be receiving. Most work on test expectancies has manipulated the test *format* (e.g., multiple-choice versus essay; McDaniel, Blishak, & Challis, 1994; Thiede, 1996) and has examined effects of expectancy manipulations on test performance itself, but not on monitoring accuracy. Less work has explored effects of anticipating different test *types* (such as memory versus inference questions) and effects on monitoring accuracy. For example, Jensen, McDaniel, Woodard, and Kummer (2014) manipulated the memory-based versus inferential nature of test items. However, they only explored effects on test performance rather than monitoring accuracy. Also, the practice test items were always on the same concepts as the target test items, which made any possible test expectancy effects confounded with testing effects. In another study, Thomas and McDaniel (2007) assessed monitoring accuracy while informing learners about the nature of the test items to expect. They distinguished between detail questions that tested for verbatim information found within single sentences of expository texts (i.e., memory questions), and conceptual questions that required information to be integrated across sentences (i.e., inference questions). However, in this study the type of test always matched the expectancy that learners were given, and what was varied was whether encoding tasks were consistent or inconsistent with the type of test (used for both expectancy and actual test questions). Because the test expectancy was not crossed with actual test type in this study, one cannot separate encoding effects from the effects of test expectancies on monitoring accuracy.

In contrast, one prior study has tested for effects of expectancy on monitoring by using a design where memory-versus-inference test expectations could either match or mismatch the tests that were given (Thiede, Wiley, & Griffin, 2011). Graduate students in education were either told to expect tests of their memory (to remember specific information) or to expect tests of their comprehension (to make connections between parts of the text). Participants read example texts, then completed example memory-test items or inference-test items matching the general description they were given. Expectancies were manipulated separately from the type of tests given for the target texts. This was done by giving all participants both memory and inference tests for the target texts, such that their expectancy was congruent with one test type but incongruent with the other test type. Thiede et al. (2011) showed a clear effect of expectancy on monitoring accuracy. Readers in the memory-expectancy condition made judgments that were significantly more predictive of memory-test performance than inference-test performance. In contrast, readers in the comprehension-expectancy condition made judgments that were more predictive of inference than memory-test performance.

There are important limitations of the Thiede et al. (2011) experiment. First, participants were graduate-level students in education with prior training about the different kinds of items that appear on tests of reading comprehension, and the different types of reading skills they are intended to measure. Such advanced students are likely to be particularly able to take advantage of test-expectancy information. Second, the lack of a baseline no-expectancy control condition prevents the conclusion that the comprehension-test expectancy improved relative metacomprehension accuracy rather than the memory-test expectancy hindering it. Third, there is no way to determine the role of the general test description versus the example test items. Participants may have picked up on some implicit differences in the example test items without noticing or tying them to the intended memory-inference distinction. Fourth, expectancies provided before reading could have impacted either the initial encoding of the target texts in a manner similar to past activity manipulations, or could have impacted the post-reading metacognitive judgment process.

Together, the following four experiments were designed to overcome these limitations and provide clearer evidence of whether typical readers are able to apply comprehension-test expectancies, adjust their metacognitive monitoring processes, and make more accurate judgments of text comprehension. After demonstrating how manipulating test expectancies can impact relative metacomprehension accuracy of undergraduates in Experiment 1, subsequent studies explored the independent contributions of providing example test items versus providing an explicit description of the test questions as requiring memory or inference (Experiment 2); the effects of establishing a test expectancy after text processing is complete (Experiment 3); and whether the benefits of test expectancies overlap and are redundant with the benefits of an activity manipulation (self-explanation) previously shown to produce large improvements in relative metacomprehension accuracy (Experiment 4).

Experiment 1

The question for Experiment 1 was whether providing readers with explicit comprehension goals and examples of inference items would improve relative metacomprehension accuracy. The main goal was to replicate the main finding from Thiede et al. (2011) and extend it to an undergraduate sample. It also sought to clarify which expectancy condition is altering default expectancies by adding a no-expectancy control condition. Including separate memory and inference tests for all target texts allowed for within-participants comparisons of whether judgments were better predictors of memory-test performance (relative metamemory accuracy) or inference-test performance (relative metacomprehension accuracy). The comprehension-expectancy condition was predicted to lead to judgments that better predicted inference-test than memory-test performance. The opposite pattern was expected in the memory-expectancy condition, and in the no-expectancy condition.

It is important to note that in all of the present studies, the example passages and example test items were *on entirely different topics* than the later target texts and tests used to assess monitoring accuracy. The example test items related to the target tests only by giving readers a general sense of the types of questions (memory versus inference) that they should expect on later tests. Furthermore, participants were not given any instruction that they should use this information about the test type to change how they read or make their test predictions. Thus, any effects of the expectancy manipulation would require that participants apply the expectancies about the general nature of the upcoming tests to modify some aspect of their metacognitive processes for the future texts.

Method

Participants. Participants were 120 undergraduates who received course credit as part of an introductory psychology subject pool. The key test of expectancy effects is contrasting metacomprehension versus metamemory accuracy levels in each of the 3 expectancy conditions. The greater accuracy for metamemory over metacomprehension accuracy in the no-expectancy condition of Thiede et al. (2011) had an estimated effect size of d = 46. A power analysis revealed that an effect this size requires 40 participants per expectancy condition (120 total) to achieve .80 power.

Design. The design was a 3 (test expectancy: none, memory, comprehension) x 2 (test type: memory, inference) mixed design. The order of the two types of tests was counterbalanced as a within-participants variable that allowed for testing how judgments differentially predicted memory versus inference test performance.

Materials. Texts and test questions are presented in Appendix A. The expository texts described complex phenomena in the natural or social sciences (antibiotics, evolution, volcanoes, intelligence tests, ice ages, monetary policy) based on materials used in prior studies (Griffin et al., 2008; Jaeger & Wiley, 2014; Thiede et al., 2011). The texts were written so that a model of the phenomenon could be constructed from the logical or causal relationships underlying each text; however, several important connections among ideas in the texts were not explicitly stated and needed to be generated by the reader. The texts varied from 650-900 words in length, had Flesch-Kincaid grade levels of 11-12, and reading ease scores in the difficult range of 31-49. For each text, one 5 item multiple-choice test was created with memory-for-detail questions, and a second 5 item multiple-choice test was created with inference questions. The distinction between memory-for-detail and inference questions is common in studies that attempt to assess understanding from expository science texts, rather than simply memory for texts (Hinze, Wiley, & Pellegrino, 2013; Karpicke & Blunt, 2011; Kintsch, 1994; Mayer; 1989; McNamara et al., 1996; Thomas & McDaniel, 2007; Wiley, Jaeger, Taylor, & Griffin, 2018).

Consistent with prior work (Jaeger & Wiley, 2014; Thiede, et al., 2011; Wiley et al., 2005), memory-for-detail questions required that the reader recognize a specific factual detail where the correct response used a highly similar surface form (words and syntax) that appeared in a single sentence of the text. For example, the test question “How many of the world's volcanoes are located on the perimeter of the Pacific Ocean?” could be answered by recalling the single text sentence “More than half of the world's volcanoes encircle the Pacific Ocean…”

In contrast, the inference questions tapped implicit relationships that could only be inferred by connecting various ideas within and across sentences and integrating them with basic world knowledge. For example, the answer to the test question “Where is the least likely place for a volcano to occur?” is not explicitly stated, but readers can infer that “C. the middle of a continent” is the best answer to this question based upon the text sentences “Volcanoes are not randomly distributed over the Earth's surface. Most are concentrated on the edges of continents, along island chains, or beneath the sea forming long mountain ranges.” The fact that the middle of a continent is, by definition, away from its “edges” is the kind of basic world knowledge that readers need to apply to understand the meaning of the words and phrases in the text, how they are related, and what they imply.

The inferences that were tested were not simple logical deductions where one could replace all concepts with abstract tokens like ‘p’ and ‘q’, and deduce the answer with certainty from the text. Rather they were inductive inferences, depending on plausible connections and integration with basic world knowledge. Such inferential connections are the crux of developing a situation-model or mental-model from expository text (Graesser & Bertus, 1998; Kintsch, 1994; Mayer, 1989). Correct answers were probabilistic, and represented the most plausible answers from among the choice options given the information in the text. Some inference items involved realizing a cause-effect relationship that was never explicitly stated in the text but was implied (for example, by mediating steps in a process). Some inference items required applying a stated claim to a new hypothetical context. Some required engaging in counterfactual reasoning to predict what would happen if a link in a causal chain were altered or removed. The incorrect inference options shared high surface overlap with the text and correct options, so they could only be rejected by inferring that their stated relationships among concepts (usually causal factors of a phenomena) were either not implied or the opposite of what was implied by the text.

In addition to being similar to test questions used in other empirical research exploring comprehension from expository texts, the inference items used here were also similar to question types used on MCAT Critical Reasoning subtest and the ACT Reading Comprehension subtest (ACTREAD) that require readers to think about inferences or implications that follow from text. The MCAT Critical Reasoning subtest uses items that range from basic memory for information mentioned in the passage, to items that ask the reader to apply information to new contexts and to consider hypothetical relations. ACTREAD includes both questions that probe for memory of verbatim information from the text as well as questions that require the reader to use reasoning skills to understand sequences of events; make comparisons; comprehend cause-effect relationships; and draw generalizations. Evidence for the validity of these tests as measures of reading comprehension comes from their correlations with standardized comprehension tests (correlation of inference items to ACTREAD, r = .35, correlation of memory-for-details items with ACTREAD, r = .30). In addition, performance on the multiple-choice inference tests used in this study correlated at *r* = .52 with performance another test of understanding (open-ended how-and-why questions for each topic) for a different sample of readers (Guerrero, & Wiley, 2018), whereas performance on open-ended comprehension questions did not correlate with performance on the multiple-choice memory-for-details tests, *r* = .06.

The number of items used for each test in this study is similar to number of items used for each passage on the standardized tests mentioned above, as well as in prior research. A recent review of all empirical research using measures of relative metacomprehension accuracy documents that 5-6 test items per passage is typical of most studies (Griffin, Mielicki, & Wiley, in press). From a practical perspective, it would be difficult to generate a larger set of unique inference items without substantially extending the length of each passage, decreasing its coherence, or assuming too much prior world knowledge on the part of the reader. This is due to the complexity inherent in each inference item which requires testing for implicit relations among ideas from the text.

In addition, the results of several norming studies support the assumed memory-versus-inference distinction between the test items (Wiley & Guerrero, in press). In the first norming study, a sample of readers were able to correctly identify the intended item type over 80% of the time (simple agreement 84.4%, ICC (1, 720) = 81.8) when given definitions for the two different item types. This is a high level of agreement compared to other work which has used only a 50% criterion for correct categorization among test item types (Nestojko, Bui, Kornell, & Bjork, 2014). In the present context, mis-categorizations were often due to negations (which were intended as requiring an inference), paraphrases (which were intended as memory-based questions), and the classification of simpler bridging inferences as memory questions (although they were intended as inference questions).

Two additional norming studies demonstrated the distinction between these item types using manipulations that affect one type but not the other. A second norming study showed that only memory-for-details questions substantially improved when texts were available during testing and respondents merely needed to search for the verbatim information that matched the correct answer. Since correct answers to inference questions are not explicitly available in the text, but rather require the reader to engage in a reasoning process, inference test performance was not substantially improved by having the texts available. This is consistent with Ferrer, Vidal-Abarca, Serrano, and Gilabert (2017) who also found that having the text available only improved performance on memory-for-details items, but not for inference items.

In a third norming study (Guerrero & Wiley, 2018), students were run individually, and were explicitly encouraged to engage in reasoning processes via the use of self-explanation prompts as they read the texts. In this condition, performance on the comprehension items substantially improved whereas performance on the memory-for-details questions did not. This is consistent with Jaeger and Wiley (2014) who also found that performance on memory-for-details items did not improve with self-explanation instructions.

Overall readers were more likely to get memory-for-details questions correct than inference questions, but importantly, performance on neither item type was at floor or ceiling. This meant there was room for readers to vary in their performance on these tests. Although the inference test items may be more difficult, and may be perceived as more difficult, such differences would only have an impact on mean test performance or mean judgment magnitudes (i.e. values). The employed measure of relative accuracy was designed and recommended precisely because of its independence from factors that impact average judgment magnitudes or test performance levels (Nelson, 1984). Nevertheless, follow-up analyses including test performance measures as a covariate were conducted.

Procedure. Participants were randomly assigned to conditions. All participants were told that they would be reading a set of texts, judging their comprehension of each text, and then taking a test for each text. Prior to reading the example texts, the memory-test-expectancy group was told that they would be tested on their "memory of specific details for each text". They read the first example text, and immediately made their judgment based on the question "How many items do you think you will get correct on a 5-item test?" They were then given a 5-item test of their memory for details presented in the text. They repeated this read-judge-test process for the other two example texts. Thus, in this study, test expectancy was manipulated using both instructions that informed participants about the nature of the tests and example test items. Following the example texts, readers were informed that the texts they had just read were for practice and now they would read and make judgments for each of the 6 critical texts. After making the judgment for the sixth critical text, they completed the first set of tests. Text order was held constant, and tests were presented in the same topic order as the texts. For the critical texts, readers completed both the memory-for-details test and the inference test for each text. Items of each test type were presented in separate blocks, counter balanced to control for order effects. There were no effects of test order, so order was collapsed for all the reported analyses. All experiments were run under an approved Institutional Review Board protocol.

The comprehension-test-expectancy group completed a similar procedure, only the general test type description and example tests were changed. Participants were told that they would be tested on their comprehension for each text and "their ability to make connections across different parts of the text". They were given inference-based questions on the 3 example texts. The no-expectancy group received the instruction that they would be "taking a test" for the critical texts and did not receive any example test items for the example texts. They did read and judge the example texts so that the procedure remained as similar as possible to other conditions.

Results and Discussion

Judgments and test performance. The primary focus of this investigation is on relative monitoring accuracy; however, as monitoring accuracy is computed as the relationship between metacognitive judgments and test performance, descriptive data are first reported for these measures. Table 1 shows the average judgments, memory test scores, and inference test scores. A one-way ANOVA revealed a marginal effect for expectancy condition on the average magnitude (i.e. value) of judgments, $F(2, 117) = 2.71$, $MSE = .62$, $p = .07$, $\eta_p^2 = .04$. Average judgment magnitude was greater for readers who received no expectancy instructions than for the other groups, but the memory and comprehension-test expectancy groups did not differ. A 3x2 (expectancy condition x test type) repeated-measures ANOVA on test performance revealed a main effect for test type with better performance on memory tests than inference tests, $F(1, 117) = 5.74$, $MSE = .01$, $p < .02$, $\eta_p^2 = .05$. The main effect for expectancy condition was marginal, $F(1, 117) = 2.48$, $MSE = .02$, $p = .09$, $\eta_p^2 = .04$. Comprehension-test expectancy tended to produce better performance on both test types. The interaction was not significant, F<1. More critical for the subsequent analyses on relative monitoring accuracy (computed as the covariance between judgments and performance) similar variance in judgments and test performance was seen across conditions, and there were no ceiling or floor effects. This was true for all experiments.

Monitoring accuracy. Relative monitoring accuracy was computed using intra-individual Pearson correlations of each participant's judgments with their corresponding test performances. (The same pattern of results was observed using Gamma correlations, as reported in Appendix B.) The statistical analyses are reported using Pearson for several reasons (see Griffin, et al., in press, for a more complete argument). When judgments and test performance are measured on non-dichotomous scales, Gamma ignores all information about the variance in the magnitude of concordances and discordances (Benjamin & Diaz, 2008; Griffin et al., in press; Schwartz & Metcalfe, 1994). In addition to the loss of statistical power due to eliminating much of the variance in the computed scores, Gamma can lead to abnormal distributions with many scores at ceiling (c.f. Wiley, Jaeger, Taylor, & Griffin, 2018). Also, participants were asked to make judgments that consisted of predicting the objective number of future test items correct, rather than subjective confidence judgments. This reduces the oft-cited problem of assuming linearity with subjective confidence judgments on a Likert-scale (Nelson, 1984).

Figure 1 shows the mean relative accuracy of judgments. A 3x2 (expectancy condition x test type) repeated-measures ANOVA revealed main effects for both expectancy condition, $F(2, 117) = 3.93$, $MSE = .15$, $p < .02$, $\eta_p^2 = .02$, and test type, $F(1, 117) = 5.23$, $MSE = .09$, $p < .03$, $\eta_p^2 = .06$. However, these effects were qualified by a significant expectancy x test type interaction, $F(2, 117) = 14.80$, $MSE = .09$, $p < .0001$, $\eta_p^2 = .20$, where expectations selectively improved monitoring accuracy in expectancy-congruent test conditions. Planned comparisons revealed that relative metamemory accuracy was greater than relative metacomprehension accuracy in the no-expectancy condition, $t(39) = 2.10$, $p < .05$, $d = .39$. This benefit for metamemory was even larger in the memory-test-expectancy condition, $t(39) = 4.23$, $p < .001$, $d = .89$ (as per Dunlop, Cortina, Vaslow, & Burke, 1996, Cohen's *d* was computed using pooled SD, for both between and within-participant simple effects). In contrast, the comprehension-test-expectancy condition showed the opposite pattern with metacomprehension being more accurate than metamemory, *t*(39) = 3.53, *p* < .001, *d* = .69.

Note that relative accuracy is statistically independent from any effects on average test performance itself (Nelson, 1984), or effects on average judgment magnitude (Griffin et al., 2013). Thus, the expectancy effects on test performance reported above cannot account for effects on relative accuracy. Correspondingly, entering both memory and inference test performance as covariates did not alter the results for any analyses.

These results show that the test-expectancy manipulation led to judgments that were more predictive of actual performance on expectancy-congruent tests. The results of the no-expectancy condition are consistent with prior research showing that readers default to memory-based rather than comprehension-based cues (Thiede et al., 2010). The memory-test expectancy reinforced this default tendency. The comprehension-test expectancy reversed this pattern, leading to higher relative metacomprehension accuracy than other conditions, and higher than the typically observed levels of .27. These results suggest that the manipulation established different test expectancies in a sample of undergraduate readers. The question pursued in the next experiment is whether the "expectancy" effects observed in Experiment 1 were due to the example test items or the explicit instruction that revealed the general nature of the tests.

Experiment 2

A combination of explicit instructions and example test items established expectancies in readers in Experiment 1. From this design, it is unknown whether the general test description would be sufficient to create this expectation, or whether example items are crucial to illustrate the type of test, or allowed readers to pick up on some other feature of the example test items. In Experiment 2, readers either got the explicit instruction that informed them about the nature of the upcoming tests, or they received the example test items. No participants in Experiment 2 received both manipulations, and the no-expectancy condition was not included. All other aspects of the procedure were the same as Experiment 1. The memory-versus-inference distinction among example items with the same multiple-choice format could be difficult for participants to notice on their own. Therefore, if the effect in Experiment 1 stemmed from this distinction rather than some other feature of the test items, then the explicit description of this distinction should have some effect.

Method

Participants. The participants were 80 undergraduates who received course credit as part of an introductory psychology subject pool. Based on effect sizes observed in Experiment 1 (.69-.89), a power analysis revealed that 20 subjects per condition would provide an 80% chance of detecting differences in monitoring accuracy due to instructional factors.

Design. The design was a 2 (manipulation type: test description or example test) x 2 (test expectancy: memory, comprehension) x 2 (target test type: memory, inference) mixed design. The order of the two types of tests was counterbalanced.

Materials and procedure. All participants in the example-test conditions were told that they would be tested after reading, but were not told the nature of the tests. These participants then read the same example texts and took the same example tests as in Experiment 1. All participants in the test-description conditions read the same brief test description used in Experiment 1 about the general nature of the upcoming tests as requiring either memory-for-details or comprehension-based connections among ideas. They were then given example texts to read, but no example test items. All other methods followed the procedure from Experiment 1.

Results and Discussion

Judgments and test performance. Table 1 shows the average judgments, memory test scores, and inference test scores. A 2x2 ANOVA on judgments revealed a significant effect of manipulation type, $F(1, 76) = 6.73$, $MSE = .67$, $p < .02$, $\eta_p^2 = .08$, an effect for expectancy condition, $F(1, 76) = 4.39$, $MSE = .67$, $p < .05$, $\eta_p^2 = .06$, and a significant interaction, $F(1, 76) = 6.04$, $MSE = .67$, $p < .02$, $\eta_p^2 = .07$. Follow-ups revealed that the exposure to the different types of example test items did not impact judgment magnitude, $t<1$, *ns*. However, participants gave higher judgments when tests were described as requiring memory than described as requiring comprehension, $t(38) = 3.69$, $p < .001$, $d = .85$.

A 2x2x2 (manipulation type x test expectancy x test type) repeated-measures ANOVA on test performance revealed a main effect for test type with better performance on memory tests than inference tests, $F(1, 76) = 10.96$, $MSE = .01$, $p < .001$, $\eta_p^2 = .14$. There was a marginal but non-significant effect of manipulation type with test performance tending to be higher for the test description condition than the example test condition, $F(1, 76) = 3.66$, $MSE = .02$, $p = .06$, $\eta_p^2 = .05$. No other effects on test performance were significant, *F*s<1.

Monitoring accuracy. Figure 2 shows the mean relative accuracy of judgments. A 2x2x2 repeated-measures ANOVA revealed main effects only for critical test type with relative metamemory accuracy being higher than relative metacomprehension accuracy, $F(1, 76) = 5.34$, $MSE = .11$, $p < .05$, $\eta_p^2 = .07$. However, this was qualified by a significant three-way interaction, $F(1, 76) = 4.10$, $MSE = .11$, $p < .05$, $\eta_p^2 = .05$. The left side of Figure 2 shows that when expectancies were manipulated via example tests alone, relative metamemory accuracy was greater than relative metacomprehension accuracy across both conditions, $F(1, 38) = 12.00$, $MSE = .11$, $p < .01$, $\eta_p^2 = .24$. There was no main effect of the example tests and no significant interaction with critical test type. In contrast, the right side of Figure 2 shows that when expectancies were manipulated via descriptions of the nature of the tests, there was the same congruency-driven expectancy x test type interaction observed in Experiment 1, $F(1, 38) = 10.21$, $MSE = .11$, $p < .01$, $\eta_p^2 = .21$. Planned comparisons revealed that metamemory was significantly higher than metacomprehension when the tests were described as memory tests, $t(19) = 2.21$, $p < .05$, $d = .69$. In contrast, metacomprehension was significantly higher than metamemory when tests were described as comprehension tests, $t(19) = 2.31$, $p < .05$, $d = .70$.

These results suggest that example inference test items by themselves did not make readers shift to comprehension cues for their judgments, but an explicit instruction about the nature of upcoming tests did. Both example test conditions showed the same bias favoring judgments that predicted memory rather than comprehension observed in the no-expectancy and memory-test-expectancy conditions in Experiment 1. This reinforces the idea that most students have a default expectancy for memory tests. In the absence of an explicit description about the nature of the test items, readers may have perceived example inference items merely as difficult memory items, thus, failing to adjust their default assumption. Based on their typical classroom experiences, participants may assume that multiple-choice format test items tend to assess verbatim memory for text details. These findings from the first two experiments are consistent with the suggestion that since readers are being primarily exposed to memory tests throughout their years of schooling (e.g., Thiede, Redford, Wiley, & Griffin, 2012), they may neither expect inference tests nor recognize them as such when given examples. Likewise, the lack of an effect for example items alone suggest that the results of Experiment 1 were not due to some other unintended idiosyncratic differences between the example test items. It appears that readers need to be given an explicit expectation about the nature of the upcoming tests in order to make more accurate metacomprehension judgments. The results do not rule out the possibility that example inference test items may enhance the comprehension-test expectancy when combined with an explicit description. To err on the side of establishing the strongest expectancy, we used the combined description-plus-examples manipulation for the subsequent two studies.

Although the results support readers playing a more active role by applying (without explicit prompting) the general information of expected test type to future metacognitive processing, it is uncertain whether this expectancy is having its effect directly upon judgment processes or by prompting readers to alter their text processing similar to the experimenter-required tasks of prior manipulations (e.g., self-explaining). Thus, the third experiment was designed to evaluate these alternative accounts.

Experiment 3

If test expectancies are altering the way that readers are encoding the texts, then test-congruent improvements in monitoring accuracy can be explained by transfer-appropriate monitoring (TAM—Dunlosky & Nelson, 1997). TAM posits that the accuracy of metacognitive monitoring will vary as a function of the match between processes engaged in prior to judgment and processes required on the test (Dunlosky, Rawson, & Middleton, 2005). Although some studies show improvements that are not adequately explained by TAM (Dunlosky & Nelson, 1997; Dunlosky et al., 2005; Weaver & Kelemen, 2003), there has been some support for TAM (e.g., Begg, Duft, Lalonde, Melinick, & Sanvito, 1989; Glenberg et al., 1987; Maki & Serra, 1992). Thomas and McDaniel (2007) have extended TAM to interpret their findings of improved monitoring accuracy due to congruence between type of encoding during reading and type of test. However, prior studies supporting TAM differ from the current ones in that they directly altered study behaviors and the processing of target information via different required experimental tasks. Thus, those findings are insufficient to predict that TAM effects might be seen in the current paradigm. Such effects are only expected if one also assumes that readers could and would apply general knowledge of test type to self-initiate changes to their text processing in the particular ways that would improve judgment accuracy.

To eliminate any effects that could result from the test expectancy being used to alter processing during encoding, the test expectancy manipulation in this experiment was introduced only *after* participants finished reading the target texts, but before predicting test performance (similar to title-before vs. title-after manipulations used by Anderson & Pichert, 1978, and Bransford & Johnson, 1972). If expectancies are only altering text processing which then happens to impact the cues available at judgment, then introducing expectancies after reading should not improve monitoring accuracy. However, if expectancies directly impact which type of cues that readers select to use at the time of judgment, then introducing post-reading expectancies should still impact judgment accuracy.

Method

Participants. The participants were 72 undergraduates who received course credit as part of an introductory psychology subject pool. Based on effect sizes observed in Experiment 2 (.69-.70), a power analysis revealed that 24 subjects per condition would provide an 80% chance of detecting differences in monitoring accuracy due to instructional factors.

Design. The design was a 3 (test expectancy *after*: none, memory, comprehension) x 2 (test type: memory, inference) mixed design. Test type order was counterbalanced.

Materials and procedure. The test expectancy manipulation used in Experiment 3 was the combined manipulation from Experiment 1, where readers received both the explicit description about the nature of the upcoming tests and the example test items. The difference between Experiment 3 and Experiment 1 was the post-reading placement of the expectancy manipulation. This meant the expectancy manipulation came after reading all the texts and before making judgments. Thiede et al. (2005) showed that merely delaying judgments does not improve relative metacomprehension accuracy unless readers also engage in a generation task that requires accessing and using the representation of the to-be-judged texts (e.g., summarizing). Thus, this change was not expected to affect overall accuracy compared to the prior experiments, and any such delay effect would be similar across conditions.

Results and Discussion

Judgments and test performance. Table 1 shows the average judgments, memory test scores and inference test scores. A one-way ANOVA on judgments revealed no significant effect for expectancy, $F(1, 69) = 1.97$, $MSE = .73$, $p = .15$, $\eta_p^2 = .05$. A 3x2 (expectancy x test type) repeated-measures ANOVA on test performance revealed a main effect for test type with better performance on memory tests than inference tests, $F(1, 69) = 11.59$, $MSE = .01$, $p < .01$, $\eta_p^2 = .16$. There was no main effect for expectancy or any interaction, *F*s < 1.23.

Monitoring accuracy. Figure 3 shows the mean relative accuracy of judgments. A 3x2 repeated-measures ANOVA revealed no main effects, *F*s<1, but there was a significant interaction, $F\ (2, 69) = 4.34$, $MSE = .12$, $p < .02$, $\eta_p^2 = .11$. Planned comparisons revealed that relative metacomprehension accuracy was better than relative metamemory accuracy in the post-reading comprehension-test-expectancy condition, $t(23) = 2.12$, $p < .05$, $d = .57$. The no-expectancy and post-reading memory-test-expectancy conditions showed an opposite trend favoring metamemory over metacomprehension accuracy, but the tests of simple effects were non-significant, *t*s(23) = 1.51 and 1.10, *p*s > .15, *d*s = .38 and .26, respectively.

Although the bias favoring metamemory in the memory-test-expectancy condition was not as strong as in Experiment 1, the same congruency-dependent interaction still emerged as well as the bias favoring metacomprehension in the comprehension-test-expectancy condition. Yet, in this experiment there was no opportunity for test expectancy to impact the reading process or initial encoding. This suggests that test expectancies are having an effect by altering how participants utilize the cues available to them at the time of judgment, rather than by solely altering encoding.

Experiment 4

The goal of creating a test-expectancy was to improve relative metacomprehension accuracy by giving readers the information they needed to actively utilize the most appropriate and diagnostic cues when making monitoring judgments. This was developed in contrast to earlier approaches where readers were explicitly directed to engage in supplemental activities during or after reading. The impact of expectancies on judgment accuracy even when established with only a post-reading manipulation suggests that expectancies are having some influence on the judgment process. Another way to evaluate whether the expectancy is having a more direct influence on the judgment process rather than indirectly via impacting text processing is to test whether introducing expectancies leads to any additional improvement in accuracy when combined with a self-explanation task manipulation that directly alters text processing. Self-explanation has already been shown to improve relative metacomprehension accuracy, arguably via increasing readers' access to situation-model-level cues (Griffin et al., 2008). If the primary benefit of the comprehension-expectancy manipulation is due to readers altering their approach to reading the texts in a way similar to self-explanation manipulations, then when combined, the effects of two manipulations should overlap, resulting in an under-additive interaction. There should be little benefit of adding a comprehension-test expectancy to a condition where readers are already engaging in self-explanation. Alternatively, if expectancies boost accuracy by altering post-reading judgment processes (as supported by Experiment 3), then there will be an additive effect (or possibly an over-additive interaction) whereby both manipulations led to significant independent improvements in relative metacomprehension accuracy.

Method

Participants. The participants were 160 undergraduates who received course credit as part of an introductory psychology subject pool. The central question was whether metacomprehension accuracy would show an under-additive interaction between getting the comprehension expectancy and engaging in self-explanation. A power analysis assuming a medium effect size of Cohen's *f* = .25, revealed that 40 participants per condition (a total of 160) would achieve a power of .80.

Design. The design was a 2 (test expectancy: no expectancy, comprehension-test expectancy) x 2 (encoding manipulation: no self-explanation, self-explanation) x 2 (test type: memory, inference) mixed design. The purpose of Experiment 4 was to contrast two different mechanisms for increasing relative metacomprehension accuracy, thus the memory-test-expectancy condition was not included. The order of the two types of tests was counterbalanced.

Materials and procedure. Half of the participants were given a comprehension-test expectancy using both the explicit instruction and example tests as in Experiments 1 and 3. The other half received the no-expectancy condition. Half of each test-expectancy condition received an additional set of self-explanation instructions identical to those used in Griffin et al. (2008). Participants were told, "As you read each text, you should try to explain to yourself the meaning and relevance of each sentence or paragraph to the overall purpose of the text. Ask yourself questions like: What new information does this paragraph add? How does it relate to previous paragraphs? Does it provide important insights into the major theme of the text? Does the paragraph raise new questions in your mind? Before you move on to the next paragraph, explain to yourself what the previous paragraph meant." Participants in the self-explanation conditions were also shown a brief example paragraph from a text on a different topic along with hypothetical statements they could make to themselves. All other materials and procedures were the same as in Experiment 1.

Results and Discussion

Judgments and test performance. Table 1 shows the average judgments, memory test scores, and inference test scores. A two-way ANOVA on judgment magnitude revealed no significant effects of expectancy condition, $F(1, 156) = 2.74$, $MSE = .62$; $p = .10$, $\eta_p^2 = .02$, or of self-explanation, $F(1, 156) = 1.19$, $MSE = .62$, $p =.28$, $\eta_p^2 = .01$, nor an interaction, $F(1, 156) = 1.42$, $MSE = .62$; $p = .24$, $\eta_p^2 =.01$. A 2x2x2 repeated measures ANOVA on test performance showed only a significant effect of test type, with better performance on the memory tests than the inference tests, $F(1, 156) = 9.63$, $MSE = .01$, $p < .01$, $\eta_p^2 = .06$. Neither the effect of expectancy condition, $F(1, 156) = 1.26$, $MSE = .02$, $p = .26$, nor self-explanation, $F(1, 156) = 1.76$, $MSE = .02$, $p = .19$, nor their interaction, $F(1, 156) = 1.96$, $MSE = .02$, $p = .16$, reached significance.

Monitoring accuracy. Figure 4 shows the mean relative accuracy of judgments. A 2x2x2 repeated-measures ANOVA revealed a significant three-way interaction, $F(1, 156) = 3.99$, $MSE = .11$, $p < .05$, $\eta_p^2 = .03$. To follow-up this significant interaction, separate ANOVAs were run for metamemory and metacomprehension. No significant effects were found in the 2x2 ANOVA for relative metamemory accuracy, *F*s < 1.56, *p*s > .21. The 2x2 ANOVA for relative metacomprehension accuracy revealed significant main effects for both self-explanation, $F(1, 156) = 16.2$, $MSE = .10$, $p < .001$, $\eta_p^2 = .09$, and expectancy $F(1, 156) = 17.0$, $MSE = .10$, $p <$

.001, $\eta_p^2 = .10$, but no significant interaction, $F(1, 156) = 2.04$, $MSE = .10$, $p = .16$, $\eta_p^2 = .01$. This lack of a two-way interaction on metacomprehension accuracy shows that there was not an under-additive effect, and that the benefit of adding an expectancy manipulation was not lessened when added to a condition that already included self-explanation. Planned comparisons for relative metacomprehension accuracy revealed that the comprehension-test-expectancy alone ($t(78) = 3.50$, $p < .01$, $d = .79$) and self-explanation alone ($t(78) = 3.34$, $p < .01$, $d = .75$) conditions resulted in greater metacomprehension accuracy than the control (no comprehension-test-expectancy, no self-explanation) condition. Further, the combined condition (that received both self-explanation and comprehension-test-expectancy) had greater relative metacomprehension accuracy than when the self-explanation manipulation was implemented alone, $t(78) = 2.22$, $p < .03$, $d = .52$. The combined condition was also superior to the comprehension-test-expectancy alone condition, $t(78) = 2.24$, $p < .03$, $d = .52$.

Adding a comprehension-test expectancy led to similar increases in relative metacomprehension accuracy even in the context of a self-explanation instruction. If the expectancy was improving metacomprehension mainly by influencing text processing in ways similar to what self-explanation does directly, then the added effect of the comprehension-test expectancy should have been significantly lessened when combined with self-explanation and compared to self-explanation by itself.

General Discussion

This series of experiments showed that expectancies about the nature of an upcoming test can and do influence the accuracy of monitoring judgments. In the first experiment, the no-expectancy control condition showed that readers' default judgments better predict performance on memory tests than performance on inference tests. Establishing a memory-test expectancy only reinforced this tendency, whereas establishing a comprehension-test expectancy inverted this pattern and led to improved relative metacomprehension accuracy and worse relative metamemory accuracy. Each additional experiment replicated the beneficial effects of establishing a comprehension-test-expectancy on relative metacomprehension accuracy compared to either no-expectancy or memory-test-expectancy conditions. Further, each subsequent experiment added new information to clarify the nature of this effect and its possible underlying mechanism. In Experiment 2, example inference items had no impact when presented without a general description of the type of test; whereas relative metacomprehension accuracy was improved by simply telling students to expect test questions that would assess their ability to make connections among ideas. In Experiment 3, a comprehension-test expectancy improved relative metacomprehension accuracy even when established only after processing of the to-be-learned information. The results of Experiment 4 revealed that comprehension-test expectancies and self-explanation activities provided unique non-overlapping contributions to improving relative metacomprehension accuracy.

Most studies on metacomprehension have attempted to improve accuracy by altering the processing of the to-be-learned material via supplemental activities required of the reader. Improvements in relative metacomprehension accuracy resulting from these manipulations can be accounted for as a direct consequence of performing these required supplemental tasks, without the reader altering their own metacognitive processes. Although these activity manipulations may have been effective at improving relative metacomprehension accuracy for the practiced set of materials, those interventions are unlikely to impart any general metacognitive knowledge or skills that students will be able to apply in future learning episodes. Without being made to engage in additional tasks in future contexts, readers are likely to continue to suffer from poor metacomprehension accuracy. In contrast, the test-expectancy manipulation explored in these studies required readers to apply general expectancies to modify their metacognitive processes as they monitored their learning on a new set of texts. The comprehension-test-expectancy instruction gave readers explicit information that they were able to use to improve their monitoring processes on future texts. Participants in various expectancy conditions were not instructed or required to do anything differently during or after reading the target texts. In fact, the benefit was even seen in a condition where the only manipulation was a general description of the test type provided after reading. The observation of a post-reading benefit from the comprehension-test-expectancy manipulation, in the absence of any alterations that may occur during encoding and processing of the target text information, makes the current efforts distinct from prior successful demonstrations of improvements to relative metacomprehension accuracy, and suggests that expectancies are affecting the judgment process. Further, the fact that expectancies still significantly boosted relative metacomprehension accuracy even on top of a direct manipulation of situation-model-level text processing (i.e., self-explanation) also favors the account that comprehension-test expectancies impact the judgment process over an account where expectancies simply affect text processing. The reduced magnitude of the expectancy effect when established after reading does imply that expectancies also alter something during encoding. However, this need not be a change to text encoding itself, but rather could be an increase in the selective attention that readers pay to various meta-experiences created during encoding. Research suggests that increased attention to meta-experiences during reading is important for optimal accuracy (Griffin et al., 2008).

Support for the Situation-Model-Cues Approach to Better Metacomprehension

Several results from this series of experiments provide support for the situation-model-cues approach to metacomprehension. First, the dissociations seen in relative metamemory and metacomprehension accuracy across the manipulations provide additional evidence that metacomprehension is not the same as the more often studied construct of metamemory. The fact that none of the present manipulations had positive effects on both metamemory and metacomprehension illustrates the need to view these as distinct constructs impacted by distinct factors. These results also show the utility of distinguishing among cues based in subjective experiences reflecting different levels of text representation. Cues tied to the surface representation, such as the feeling that one can recall exact words from the text, are likely to be diagnostic only for metamemory judgments. Accurate metacomprehension judgments depend on subjective experiences tied to the quality of readers' situation models, such as a sense that they understand the connection between ideas in the text.

Consistent with prior work, these results show that these readers tended to default to memory-based cues when asked to judge their understanding of text, and suggest that many students may not appreciate what text comprehension entails. These findings converge with more ecologically-valid correlational data suggesting that students who know to use comprehension rather than memory as their reading goal have more accurate metacomprehension. For example, in Thiede et al. (2010), the minority of college readers who reported basing their comprehension judgments on their ability to link ideas contained in texts were seen to have higher relative metacomprehension accuracy. Correspondingly, middle school (7th and 8th grade) students whose early literacy education focused on deep understanding and inference-building as explicit learning goals have been shown to have better relative metacomprehension accuracy and make more effective restudy choices compared to students with more typical schooling experiences (Thiede et al., 2012).

Although many readers may default to memory-based cues, this tendency was able to be altered by the introduction of a single sentence informing readers about the type of test they should expect. The fact that this subtle manipulation could have such a large impact illustrates the inherent ambiguity in what people think it means to monitor their understanding of text. This ambiguity could be a major reason why monitoring accuracy when learning from text has been so notoriously poor, and so much worse than the near-perfect monitoring accuracy observed for delayed judgments of learning for word pairs (Nelson & Dunlosky, 1991). A question like 'How well can you recall the word that was paired with DOG?' has fewer potential meanings than 'How well did you understand the passage about digestion?' Inside and outside of classrooms, learners are likely to be unclear about their goals for learning when reading complex textual explanations, and are therefore likely to be unclear about what they should be monitoring. Creating explicit comprehension expectancies allows readers to base their judgments of monitoring on a more diagnostic reference point.

When interpreting these findings it is critical to remember that relative accuracy measures were designed precisely so that overall differences in judgment magnitude or test performance would not have an impact (Nelson, 1984). So although performance on memory-test items was better than performance on the inference-test items, such a difference would not translate to differences in relative accuracy. In fact, across all four experiments the effects of test expectancy on test performance and judgment magnitude never matched the pattern of effects on relative metacomprehension accuracy. Relative accuracy depends on the ordering of a set of judgments and whether they align with a set of performances within an individual. Because a general difference in perceived difficulty would have a similar impact on the judgments for all the texts, it would not impact relative accuracy. Any viable explanation for improvements in relative accuracy must depend on something that can be applied differentially when judging each specific text or learning episode, such as using relative differences in how well readers think they understand the inferential relations in each text.

Given that the inference tests were objectively more difficult (i.e., they resulted in lower average test performance), it seems plausible that readers might have picked up on some general sense of difficulty from the example test items. However, average judgment magnitudes did not differ between conditions that received either memory or inference example test items. Rather, judgment magnitudes only between the Experiment 2 conditions that received only a general description of the tests as assessing either memory or inferences. Further, if readers were using other idiosyncratic features of the example test items as a basis for their judgments, then the same differences in relative accuracy should have emerged in the no-description, practice-test-only conditions. However, no differences were seen. Instead, providing only a general test description that referred to the intended memory-versus-inference distinction was sufficient to produce the effects on its own. This pattern of results is the opposite of what would be expected if the improvements in relative metacomprehension accuracy were due to difficulty or some unintended difference between the memory and inference-test items. Further, the results of Experiments 2 through 4 are best explained by the hypothesis that a giving readers a comprehension-test-expectancy provides them with an appropriate metacognitive goal for their monitoring processes, which helps them to align their judgments with their actual performance on comprehension tests.

This line of research focused exclusively on one particular metacognitive measure: relative metacomprehension accuracy. One primary reason for this emphasis is that only relative measures depend upon accurate online monitoring of comprehension during different individual learning episodes (Griffin, Wiley, & Salas, 2013; Griffin, Mielicki, & Wiley, in press). Other measures (calibration, absolute accuracy, and confidence bias) confound differences in monitoring processes with differences in comprehension itself (Maki, 1998; Nelson, 1984). For example, if a person is overconfident, then improving their performance with a manipulation will appear to reduce the amount of error in their judgments. Judgments based in heuristic cues, including a reader's a priori assumptions about their own ability, can predict absolute levels of performance (and may result in reductions in error on measures of calibration), but they do so without depending upon the online monitoring of different learning episodes or reflecting on meta-experiences (Flavell, 1979; Griffin, Wiley, & Salas, 2013). Online evaluation and reflection on different, specific learning episodes is required in order for a reader to make effective decisions about how to regulate their study behaviors, and how to prioritize which topics they need to study. That construct is best captured by measures of relative accuracy.

Yet, relative accuracy is but one measure that can be explored among a number of other measures of judgment-performance relationships (absolute accuracy and confidence bias, Maki, 1998; metacognitive calibration, Linderholm & Zhao, 2008; Nietfeld, Enders, & Schraw, 2006). Although these other measures do not reliably correlate with relative metacomprehension accuracy, and because absolute and relative measures are often impacted by different factors (Griffin, Jee, & Wiley, 2009, Maki, 1988; Nelson, 1984), future research should seek to identify conditions that improve both relative and absolute accuracy, which might support the most effective self-regulated learning (Dunlosky & Rawson, 2012).

One other observation about the research reported here is that these studies did not attempt to directly measure test expectancy or cue basis. Instead the readers' test expectancies and cue bases were inferred from the effects that the manipulations had on relative monitoring accuracy. Past work has attempted to assess expectancies, cue use, and judgment bases more directly by using retrospective reports (Thiede et al., 2010). Adding such methods to future studies could provide a further test of the plausibility of the current account.

Conclusions

The results across four experiments show robust and reliable benefits from generalized test expectancies that learners can apply to monitoring their comprehension of future texts. These benefits do not require exposure to example test items. Further, test expectancies seem to impact the judgment process itself (rather than just initial encoding), and improve judgment accuracy beyond directly altering the reading process such as by having readers engage in additional activities (like self-explanation) that generate appropriate meta-experiences and diagnostic judgment cues. The results highlight the theoretical importance of clarifying what is meant by "comprehension" and differentiating between mnemonic and situation-model-based cues as a basis for accurate metacomprehension when learning from explanatory, expository science texts.

An obvious limitation of prior work on metacomprehension accuracy is that most of it has been done in laboratory settings. Ultimately, more work is needed in classroom settings to be able to apply these results and make recommendations for practice. However, one recent study has shown that manipulations similar to those studied here may improve metacomprehension accuracy in an actual classroom context (Wiley et al., 2016). In this study done within a college course on Research Methods, an intervention condition received a combination of self-explanation and comprehension-test-expectancy instructions. Using a set of passages from assigned readings for the course as stimuli, this combined condition led to improved relative metacomprehension accuracy. In addition, when students were given the chance to actually re-study the reading assignments, students who were in the intervention condition used more effective study strategies, and got higher scores on classroom quizzes on these topics (Wiley et al., 2016). Taken together, this prior course-based study and the current set of results suggest that combining comprehension-test-expectancies with text-processing manipulations like self-explanation offer promise for improving self-regulated study in authentic classroom settings.

References

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*. *128,* 110-118.

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.

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 73-94). New York: Psychology Press.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11,* 717-726.

Brunswik, E. (1956). *Perception and Representative Design of Psychological Experiments.* Berkeley: University of California Press.

Dunlop, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170-177.

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., & Metcalfe, J. (2009). *Metacognition: A textbook for cognitive, aging, lifespan, & applied psychology*. London: Sage Publications.

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. *The Oxford handbook of metamemory,* 23-37.

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,* 4-49.

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,* 37-56.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22,* 271-280.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52,* 551-565.

Ferrer, A., Vidal-Abarca, E., Serrano, M. Á., & Gilabert, R. (2017). Impact of text availability and question format on reading comprehension processes. *Contemporary Educational Psychology, 51*, 404-415.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. *American Psychologist, 34*, 906–911.

Fukaya, T. (2013). Explanation generation, not explanation expectancy, improves metacomprehension accuracy. *Metacognition and Learning*, *8*, 1-18.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116,* 119-136.

Graesser, A. C., & Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading, 2,* 247-269.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*, 1001–13.

Griffin, T. D., Mielicki, M. K., & Wiley, J. (in press). Improving students' metacomprehension accuracy. To appear in J. Dunlosky & K. Rawson (Eds.), *Cambridge handbook of cognition and education*. New York, NY: Cambridge University Press.

Griffin, T. D., Wiley, J., & Salas, C. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In R. Azevedo & V. Aleven (Eds.) *International Handbook of Metacognition and Learning Technologies* (pp. 19-34). Springer Science.

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.

Guerrero, T. A. & Wiley, J. (2018). Effects of text availability and reasoning processes on test performance. To appear in *Proceedings of the 40th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69*, 151-164.

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction, 34,* 58-73.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test … or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*, 307-329.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6018), 772-775.

Kintsch, W. (1994). Text comprehension, memory, and learning. *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.

Linderholm, T., & Zhao, Q. (2008). The impact of strategy instruction and timing of estimates on low and high working-memory capacity readers' absolute monitoring accuracy. *Learning and Individual Differences, 18*, 135-143

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–144). Hillsdale, NJ: Lawrence Erlbaum.

Maki, R. H., & Serra, M. (1992). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology*, *84*, 200–210.

Mayer, R. E. (1989). Models for understanding. *Review of Educational Research, 59*, 43-64.

McDaniel, M. A., Blischak, D. M., & Challis, B. (1994). The effects of test expectancy on processing and memory of prose. *Contemporary Educational Psychology*, *19*, 230–248.

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.

Mielicki, M. K., Griffin, T. D., & Wiley J. (2017, July). *A meta-analysis of metacomprehension*. Paper presented at the 27th Annual Meeting of the Society for Text and Discourse, Philadelphia, PA.

Nelson, T. O. (1984). A comparison of current measures of feeling-of-knowing accuracy. *Psychological Bulletin*, *95*, 109–133.

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.

Nestojko, J. F., Bui, D. C., Kornell, N., & Bjork, E. L. (2014). Expecting to teach enhances learning and organization of knowledge in free recall of text passages. *Memory & Cognition, 42*, 1038-1048.

Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement, 66*, 258-271.

Otero, J.C., León, J.A., & Graesser, A.C. (Eds.). (2002). *The psychology of science text comprehension*. Mahwah, NJ: Lawrence Erlbaum.

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., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*, 1004-1010.

Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction, 22,* 262-270.

Roediger, H.L. & Karpicke, J.D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255.

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science, 6*, 132-137.

Schwartz, B.L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe and A.P. Shimamura (Eds.) *Metacognition: Knowing about knowing* (pp. 93-113). Cambridge, MA: MIT Press.

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *Quarterly Journal of Experimental Psychology A*, *49*, 901–918.

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., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed keyword effect on metacomprehension accuracy. *Journal of Experiment Psychology: Learning, Memory and Cognition*, *31*, 1267–1280.

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

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

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

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*, 264–273.

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.

Van Loon, M. H., de Bruin, A. B. H., van Gog, T., van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effect of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143 – 154.

Weaver, C. A., III (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16*, 214-222.

Weaver, C. A., & 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., III, Bryant, D. S., & Burns, K. D. (1995). Comprehension monitoring: Extensions of the Kintsch and van Dijk model. In C. A. Weaver III, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 177-193). Hillsdale, NJ: Erlbaum.

Wiley, J., Griffin, T. D., Jaeger, A. J., Jarosz, A. F., Cushen, P. J., & Thiede, K. W. (2016). Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied, 22*, 393-405.

Wiley, J., Griffin, T., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology*, *132*, 408–428.

Wiley, J. & Guerrero, T.A. (in press). Prose comprehension beyond the page. To appear in Millis, K., Magliano, J., Long, D., & Wiemer, K. (Eds). *Deep learning: Multi-disciplinary approaches*. Routledge/Taylor and Francis.

Wiley, J., Jaeger, A. J., Taylor, A. R., & Griffin, T. D. (2018). When analogies harm: The effects of analogies and valid cues on the metacomprehension of science text. *Learning & Instruction, 55*, 113-123.

Wiley, J., & Myers, J. L. (2003). Availability and accessibility of information and causal inferences from scientific text. *Discourse Processes, 36*, 109-129.