The Effects of Comprehension-Test Expectancies on Metacomprehension Accuracy
Thomas D. Griffin, Jennifer Wiley, and Keith W. Thiede

The Effects of Comprehension-Test Expectancies on Metacomprehension Accuracy

Thomas D. Griffin and Jennifer Wiley
University of Illinois at Chicago

Keith W. Thiede
Boise State University

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 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 intraindividual 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 suboptimal choices such as failing to resudy poorly understood information while attempting to learn from texts (Maki, 1998; Thiede, Anderson, & Therriault, 2003; Wiley et al., 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 empir-
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, 2016; 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 metaexperiences serve as the basis for accurate judgments, and that such experiences are generated during encoding or use of knowledge. Based on Flavell’s (1979) model, Griffin, Wiley, and Salas (2013) proposed a distinction between more appropriate (or more diagnostic) judgments, based on Flavell’s (1979) metaexperience 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 nonexperiential 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 (1998) theoretical framework 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 text-based 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, León, & 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 et al., 2000; Weaver et al., 1995; Wiley et al., 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 metaexperiences 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, Griffin, Wiley, and Anderson (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 et al., 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), rereading (Dunlosky & Rawson, 2005; Griffin et al., 2008; Rawson et al., 2000), and text unscrambling (vs. 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; Korniak, 1997), these activities lead to more accurate judgments because they require additional cognitive processing that generates metaexperiences 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 metaexperiences. 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 et al., 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 metaexperiences. Metacognitive strategies refer to metalevel goals and strategies that a learner needs to actively apply in order to either generate or make optimal use of metaexperiences for monitoring their learning. Prior manipulations have instructed readers to engage in additional activities, which directly alter readers’ cognitive actions during text process-
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, 1998; Maki & Serra, 1992), 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 metaexperiences 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. Prior work on test expectancies has manipulated the test format (e.g., multiple-choice vs. essay; McDaniel, Blischak, & 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 vs. 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, Wiley, and Griffin (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 postreading 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 test 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 re-
quiring 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 metacomprehension 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 vs. 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 three 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) × 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 test; 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 (Grassner & 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 (e.g., 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, 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 five to six 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, miscategorizations 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 (Guerrero & Wiley, in press). Because 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, in press), 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 five-item test?” They were then given a five-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 six 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 three 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
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 mon-
itoring accuracy is computed as the relationship between metacog-
nitive 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, $F(2, 117) = 2.71, MSE =
.62, p = .07, \eta^2_g = .04$. 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 $3 \times 2$ (Expectancy Condition × Test
Type) repeated-measures ANOVA on test performance revealed a
main effect for test type with better performance on memory tests
than inference tests, $F(1, 117) = 5.74, MSE = .01, p < .02, \eta^2_g =
.05$. The main effect for expectancy condition was marginal, $F(1,
117) = 2.48, MSE = .02, p = .09, \eta^2_g = .04$. 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 (com-
puted 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 intradividual Pearson correlations of each par-
ticipant’s judgments with their corresponding test performances.
(The same pattern of results was observed using Gamma correla-
tions, 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 nondichotomous 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 (cf. Wiley et al., 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 as-
suming linearity with subjective confidence judgments on a Likert-
scale (Nelson, 1984).

Figure 1 shows the mean relative accuracy of judgments. A
$3 \times 2$ (Expectancy Condition × Test Type) repeated-measures
ANOVA revealed main effects for both expectancy condition,
$F(2, 117) = 3.93, MSE = .15, p < .02, \eta^2_g = .02$; and test type,
$F(1, 117) = 5.23, MSE = .09, p < .03, \eta^2_g = .06$. However,
these effects were qualified by a significant Expectancy × Test
Type interaction, $F(2, 117) = 14.80, MSE = .09, p < .0001,
\eta^2_g = .20$, 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-

---

Table 1

<table>
<thead>
<tr>
<th>Experiment and Condition</th>
<th>Judgment (SD)</th>
<th>Memory (SD)</th>
<th>Inference (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1: Expectancy condition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No expectancy</td>
<td>2.62 (.83)</td>
<td>.50 (.11)</td>
<td>.47 (.10)</td>
</tr>
<tr>
<td>Memory</td>
<td>2.22 (.81)</td>
<td>.49 (.12)</td>
<td>.46 (.11)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>2.37 (.71)</td>
<td>.53 (.13)</td>
<td>.51 (.12)</td>
</tr>
</tbody>
</table>

| Experiment 2: Expectancy—Manipulation type | | | |
| Memory—Example tests | 2.18 (.92) | .49 (.12) | .45 (.09) |
| Comprehension—Example tests | 2.25 (.75) | .54 (.15) | .49 (.09) |
| Memory—Test description | 3.11 (.83) | .55 (.15) | .49 (.13) |
| Comprehension—Test description | 2.28 (.78) | .58 (.12) | .51 (.09) |

| Experiment 3: Postreading expectancy condition | | | |
| No expectancy | 2.78 (1.01) | .48 (.13) | .45 (.08) |
| Memory | 2.27 (.70) | .49 (.13) | .42 (.08) |
| Comprehension | 2.35 (.86) | .53 (.13) | .47 (.10) |

| Experiment 4: Expectancy—Explanation condition | | | |
| No expectancy—No explanation | 2.82 (.70) | .51 (.13) | .46 (.08) |
| Comprehension—No explanation | 2.54 (.63) | .55 (.14) | .51 (.13) |
| No expectancy—Explanation | 2.47 (.89) | .55 (.14) | .51 (.10) |
| Comprehension—Explanation | 2.48 (.90) | .53 (.13) | .53 (.12) |

*Note.* Judgments were on a 0–5 scale. Test performance is proportion correct, with chance guessing being .25.
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) \( \times \) 2 (Test Expectancy: Memory, Comprehension) \( \times \) 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 2 \( \times \) 2 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, \( t < 1 \). 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 2 \( \times \) 2 \( \times \) 2 (Manipulation Type \( \times \) Test Expectancy \( \times \) 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 nonsignificant 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, \( Fs < 1 \).
Monitoring accuracy. Figure 2 shows the mean relative accuracy of judgments. A $2 \times 2 \times 2$ repeated-measures ANOVA revealed main effects only for critical test type with relative metacognition accuracy being higher than relative metacomprehension accuracy, $F(1, 76) = 5.34$, $MSE = .11$, $p < .05$, $\eta^2_g = .07$. However, this was qualified by a significant three-way interaction, $F(1, 76) = 4.10$, $MSE = .11$, $p < .05$, $\eta^2_g = .05$. The left side of Figure 2 shows that when expectancies were manipulated via example tests alone, relative metacognition accuracy was greater than relative metacomprehension accuracy across both conditions, $F(1, 38) = 12.00$, $MSE = .11$, $p < .01$, $\eta^2_g = .24$. 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, $F(1, 38) = 10.21$, $MSE = .11$, $p < .01$, $\eta^2_g = .21$. Planned comparisons revealed that metacognition was significantly higher than metacomprehension when the tests were described as memory tests, $t(19) = 2.21$, $p < .05$, $d = .69$. In contrast, metacollection was significantly higher than metacomprehension 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 the 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 because 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 metacognition 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, Melnick, & 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 postreading 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) × 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 postreading placement of the expectancy manipulation. This meant the expectancy manipulation came after reading all the texts and before making judgments. Thiede, Dunlosky, Griffin, and Wiley, (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 with 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, η² = .05. A 3 × 2 (Expectancy × 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, η² = .16. There was no main effect for expectancy or any interaction, F(3, 21) < 1.23.

Monitoring accuracy. Figure 3 shows the mean relative accuracy of judgments. A 3 × 2 repeated-measures ANOVA revealed no main effects, Fs < 1, but there was a significant interaction, F(2, 69) = 4.34, MSE = .12, p < .02, η² = .11. Planned comparisons revealed that relative metacomprehension accuracy was better than relative metamemory accuracy in the postreading comprehension-test-expectancy condition. t(23) = 2.12, p < .05, d = .57. The no-expectancy and postreading memory-test-expectancy conditions showed an opposite trend favoring metamemory over metacomprehension accuracy, but the tests of simple effects were nonsignificant, ts(23) = 1.51 and 1.10, ps > .15, ds = .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 postreading 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 underadditive 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 postreading judgment processes (as supported by Experiment 3), then there will be an additive effect (or possibly an overadditive 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 sub-
ject pool. The central question was whether metacomprehension accuracy would show an underadditive 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) $\times$ 2 (Encoding Manipulation: No Self-Explanation, Self-Explanation) $\times$ 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, Wiley, and Thiede (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 2 $\times$ 2 $\times$ 2 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 2 $\times$ 2 $\times$ 2 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 metacomprehension and metacomprehension. No significant effects were found in the 2 $\times$ 2 ANOVA for relative metacomprehension accuracy, $F(1, 156) < 1.56, p > .21$. The 2 $\times$ 2 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 showed that there was not an underadditive 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 expectations 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 metacomprehension 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 selfExplanation activities provided unique nonoverlapping 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 those 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 postreading 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., selfExplanation) 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 metaexperiences created during encoding. Research suggests that increased attention to metaexperiences 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 readers tend 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 (seventh and eighth 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 with 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.

Limitations and Future Directions

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 differed 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 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 et al., in press; Griffin et al., 2013). 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 on 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 metacriteria (Flavell, 1979; Griffin et al., 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, 1998; 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 metaexperiences 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 restate 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


TEST EXPECTANCIES


(Appendices follow)
The first five items for each topic are memory-for-details questions and the last five items are inference questions. Correct answers are indicated with “***”.

### Appendix A

**Texts and Test Questions**

#### Volcanoes

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

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

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

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

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

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

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

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

1. How many of the world’s volcanoes are located on the perimeter of the Pacific Ocean?
   A. None
   B. About a third
   C. Over half
   D. Almost all

2. What is true of shield volcanoes?
   A. They have steep sides.
   B. They are located near oceans.
   C. They erupt violently.
   D. They are also known as stratovolcanoes.

3. What magnitude earthquake accompanied the Mt. St. Helens eruption?
   A. 2.3
   B. 4.2
   C. 5.1
   D. 7.2

4. How many plates make up the earth’s crust?
   A. 2
   B. 7
   C. 12
   D. About 20

5. What is true of andesitic magma?
   A. It contains low amounts of silica.
   B. It contains low amounts of sulfur.
   C. It contains high amounts of magnesium.
   D. It contains high amounts of gas.

6. Where is the least likely place for a volcano?
   A. In the middle of a continent
   B. At the edge of an ocean
   C. On islands
   D. Under the ocean

7. What happens where plates diverge?
   A. Continental crust is recycled
   B. Earthquakes
   C. Violent eruptions
   D. New crust is formed

8. Which is true of converging oceanic and continental plates?
   A. The oceanic plate is pushed deep into the mantle.
   B. They are generally free of earthquakes.
   C. Continental plates are denser than oceanic plates.
   D. The two plates push up on each other and form mountains.

9. What causes violent volcanic eruptions?
   A. Fluid magmas that are low in silica
   B. Magmas that come from melted continental plates
   C. Magmas that are high in basalt
   D. Magmas that come from melted oceanic plates

(Appendices continue)
10. Which does NOT cause the creation of volcanoes?
   A. oceanic-continental plate convergence
   B. oceanic-oceanic plate convergence
   C. continental-continental plate convergence
   D. magma rising to the earth’s surface

Evolution

The first life forms on Earth were simple one-celled organisms. They appeared almost 4 billion years ago. Over time, these simple, bacteria-like cells evolved into the increasingly complex, multicelled organisms that we see today. The earliest ape-like primates, the ancestors of modern humans, monkeys, and apes, began to evolve from other mammals less than 100 million years ago. Several humanlike primate species began appearing about 5 million years ago. Modern human beings are the only humanlike primate species that still exists today.

Individuals within a species are not identical, and can differ in many ways. During sexual reproduction, genes of the parents are sorted and recombined, producing a great variety of gene combinations in their offspring. These different genetic combinations produce small differences in physical and behavioral traits of the offspring. Genetic variation causes some offspring to run faster than others, different coloring, different size, or different responses to predators. The new genetic combinations produced by sexual reproduction are just one source of genetic variation.

Although genetic instructions may be passed down virtually unchanged for many thousands of generations, occasionally some of the genetic information in a cell is altered. Deletions, insertions, or substitutions of genetic information occur due to random errors in copying, or because of environmental stressors like radiation. These mutated versions of a parent’s genes can be passed on to offspring resulting in physical or behavioral traits that other members of a species do not possess.

Although genetic variations are random, which variations survive and which die out is not. Genetic variations, known as “mutations,” randomly produce new characteristics that may help, harm, or have no effect on an organism’s ability to satisfy its basic needs and have healthy offspring. Most species produce an abundance of offspring, and not all of them will survive to adulthood and have offspring of their own. Some mutations cause diseases or deformities that lead to death or make survival unlikely. Also, limited food, space, and other resources in the environment mean that organisms compete with each other for survival. The organisms with the genes that are best suited to survival and reproduction in their particular environment are the most likely to pass on their genes to future generations.

These minor variations in genetic traits occur again and again over thousands of generations and each of the small variations add up. After many generations, organisms may possess so many different physical and behavioral traits that they seem almost nothing like their very distant ancestors. Eventually an enormous variety of organisms will exist that share common ancestors. While these organisms differ in numerous important ways, they share some basic inherited features. For example, all vertebrates, from fish to humans, have a common basic body structure characterized by a segmented body and a hollow main nerve cord along the back.

The fossil record reveals profound changes in the kinds of living things that have inhabited our planet over its long history. More than 99% of the species that have ever lived on the earth are now extinct and the oldest fossils of modern creatures like humans are still extremely recent compared with the age of life on Earth. Fossils reveal diverse creatures that have come into and out of existence. An organism’s remains are rarely fossilized, so some small species of the past may have no remains buried in the Earth. However, scientists have been lucky enough to find just some of the fossils of transitional creatures who link modern species to their ancient ancestors.

1. When did the first life forms appear?
   A. about 4 billion years ago
   B. about 5 billion years ago
   C. about 7 billion years ago
   D. about 100 billion years ago

2. When did human-like primate species begin appearing?
   A. 5 thousand years ago
   B. 50 thousand years ago
   C. 5 million years ago
   D. 50 million years ago

3. What percentage of the species that have lived on Earth are now extinct?
   A. 20
   B. 50
   C. 80
   D. 99

(Appendices continue)
4. Which of the following best describes the first life forms on Earth?
   A. simple one-celled organisms
   B. the result of errors in genetic copying
   C. the result of natural selection
   D. unrelated to modern creatures

5. What can cause deletions, insertions, or substitutions of genetic information to occur?
   A. an organism’s desire to be more adapted to their environment
   B. random errors in copying
   C. the sharing of common ancestors
   D. the fact that most genes are beneficial for survival

6. Which of the following best describes the process of evolution?
   A. It is completely random.
   B. It leads to the extinction of some species.
   C. It only began in the last million years.
   D. Few species have gone extinct over time.

7. What would happen if all organisms since the beginning of life on Earth only produced offspring that were genetically identical to themselves?
   A. Evolution could not occur.
   B. Evolution could still occur due to natural selection.
   C. Evolution would take a lot longer.
   D. Organisms would only produce the number of offspring the environment could support.

8. Which of the following represents a change that would make it much less plausible that humans could have evolved from simple life forms?
   A. if each genetic mutation produced only minor variations
   B. if life on Earth began only 10,000 years ago
   C. if other human-like primate species existed
   D. if only a few fossils of species that link humans to ancient creatures could be found

9. What would happen if the Earth’s temperature suddenly dropped by 70 degrees?
   A. Some current species would go extinct.
   B. The amount of genetic variation would be drastically increased.
   C. Most animals without any fur would grow fur to be more adaptive.
   D. Individuals who are least suited to cold weather be most likely to pass on their genes.

10. If an organism learns from its experience where to best find food, which of the following is likely to occur?
    A. It will survive better and be more likely to pass on its genes.
    B. Its offspring will be born with that knowledge and won’t have to learn it.
    C. All organisms will change their genes to be more like that one.
    D. Both A and B.

**Ice Ages**

An ice age is a period of time—usually millions or tens of millions of years—when vast glaciers cover as much as a third of the Earth’s land surface. Average global temperatures can drop by as many as 12 degrees Celsius overall. The latest Ice Age began about 2.5 million years ago, and ended approximately 15,000 years ago. Average global temperatures decreased by approximately 8 degrees Celsius. Sea-level was lowered substantially due to the amount of water that was frozen in the glaciers. Ice core analysis indicated there were reduced amounts of carbon dioxide in the atmosphere. Giant ice sheets that originated at the North Pole advanced and retreated many times in North America and Europe. The movement of the glaciers coincided with cycles of warm and cold periods in the Earth’s temperature. Throughout history, cycles of changes in global temperatures usually occur every 100,000 years or so. Each cycle consists of a long, generally cold period during which the entire Earth cools, followed by a relatively short warm period during which Earth warms up rapidly.

(Appendices continue)
We are now in a warming period that has lasted more than 10,000 years, which is longer than many of the previous warming intervals. Warm temperatures over the last century have been attributed to the increased man-made release of carbon dioxide (CO2). CO2 prevents long-wave radiation from escaping from the Earth into space. The more CO2 there is in the atmosphere, the more long-wave radiation is kept from leaving the Earth. The more radiation that is trapped, the hotter the Earth becomes. This trapping of radiation works like a gardener’s greenhouse, and this phenomenon is commonly known as the “Greenhouse Effect.”

CO2 is a common gas that is contained in the Earth’s atmosphere. CO2 is released whenever organic matter decays, and when carbohydrates are broken down by plants and animals in the process of respiration. The burning of fossil fuels also releases large amounts of CO2.

Carbon dioxide can be removed from the atmosphere also. CO2 can be combined with other minerals in the ground and buried, or absorbed into the oceans or trapped in ice and snow. Green plants also absorb carbon dioxide from the atmosphere, and through the process of photosynthesis, form carbohydrates.

The release and storage of CO2 is a natural process and proceeds in a circular fashion. For example, plants convert CO2 from the atmosphere into carbohydrates, which they use to grow. When the plant dies, the carbohydrates that the plant made are converted back into CO2 through the decaying process. At any time it is possible for there to be more CO2 being stored than released, and also vice-versa. Thus, the amount of CO2 in the atmosphere can fluctuate.

The amount of radiation that the Earth receives from the sun can also fluctuate. Fluctuations in solar radiation can change average global temperatures by up to 4-6 degrees Celsius. The amount of solar radiation that the sun emits can vary. For example, an increase in the amount of sunspots on the Sun’s surface has been correlated with an increase in the amount of energy that is output by the Sun. The amount of solar radiation energy that actually reaches the Earth is influenced by the distance the Sun’s rays must travel to reach the Earth, and also the angle at which the Sun’s rays strike the surface of the Earth. The farther that light rays travel, the less energy will be contained in the Sun’s rays. Cyclical changes in the shape of the Earth’s orbit around the Sun influence how far the Sun’s rays have to travel. When the Earth’s orbit is extremely oval-shaped, the distance from the Earth to the Sun can vary greatly. The farther the Earth from the Sun, the less solar radiation reaches the Earth.

Other cyclical changes in the tilt of the Earth’s axis vary the angle at which light energy strikes the surface of the Earth in a given region. If the Sun’s rays strike the Earth at a great angle, for example as it does at the North Pole, solar energy is reflected off of the Earth, rather than being absorbed into it. When light strikes a region at a great angle, not very much of the solar radiation is absorbed by the Earth. When a region receives less solar radiation, there is less energy to warm that area. Less heat energy leads to cooler temperatures. Cooler temperatures can cause more snow and ice to form. Snow and ice can reflect what little solar energy reaches the surface of the Earth back into space. The formation of snow and ice can also steal large amounts of CO2 from the atmosphere and trap it in a frozen, solid form.

Through the course of millions of years, the surface of the earth also changes. Continents collide and split apart, mountains are uplifted and eroded, volcanoes erupt, and ocean basins open and close. These changes alter the size and elevation of the continental land masses. Different elevations of land masses support different types of climates. Land masses at high elevations usually support colder climates. For example, the tops of mountains high above sea-level are usually covered in snow and plains at sea-level are usually warm. These events also release minerals in the Earth’s crust. These minerals are often carried by rivers to the sea, where they can be absorbed into the atmosphere. In this way, CO2 and other compounds can be released from their solid mineral forms and introduced into the atmosphere.

Changes in geography also affect the ocean by the opening and closing of gateways that carry currents. A change in ocean currents affects how water flows from one area of the Earth to another. A majority of Earth’s heat energy is transferred around the globe by the ocean currents. More heat energy is stored in the oceans than in the atmosphere. Surface ocean currents assist in the transfer of heat from low to high latitudes.

The Earth might be due for another Ice Age. However, not all scientists are convinced that there will be one. Some believe that the man-made release of CO2 into the atmosphere might prevent the Earth from cooling sufficiently. On the other hand, some scientists believe that the recent global warming might actually increase the magnitude of the cooling period, rather than prevent it.

1. How much of the Earth is covered by glaciers during an ice age?
   A. less than 10%
   B. about a third
   C. over half
   D. almost all

2. How long has the current warming period lasted?
   A. 2.5 million years
   B. 100,000 years
   C. 50,000 years
   D. 10,000 years

(Appendices continue)
3. What is NOT true of CO2?

A. It is a common gas in the earth’s atmosphere.
B. It is released by decaying plants.
C. It is released by burning fossil fuels.
D. It cannot be removed from the atmosphere.∗

4. How much do average global temperatures lower during an ice age?

A. 4–6 degrees Celsius
B. 8–12 degrees Celsius∗
C. 20–25 degrees Celsius
D. 40–50 degrees Celsius

5. What is the greenhouse effect?

A. the absorption of CO2 by growing plants
B. the trapping of radiation due to CO2∗
C. the increase in heat of the earth due to sunspots
D. the increase in burning of fossil fuels

6. What is true about ice ages?

A. Regional temperatures within an ice age do not fluctuate.
B. Sea levels are higher during an ice age.
C. Northern regions of the Earth are covered with ice∗.
D. Ice ages occur because the temperature of the core of the Earth cools.

7. What is true of the oceans?

A. Changes in ocean currents could cause glaciers to form or retreat.∗
B. Ocean currents follow the same path around the globe as they did 2.5 million years ago.
C. More heat energy is stored in the atmosphere than the oceans.
D. The oceans keep a constant temperature.

8. Higher levels of CO2 in Earth’s atmosphere lead to . . .

A. higher sea levels∗
B. the creation of mountain ranges
C. the formation of more ice and snow
D. changes in the Earth’s axis

9. What can cause less solar radiation to be absorbed by the Earth?

A. when the Earth’s orbit is closer to the Sun
B. sunspots
C. the formation of more mountain ranges∗
D. the seasons

10. What is true of Earth’s temperature?

A. It goes through long warming cycles followed by short cooling cycles.
B. Temperature changes are random and unpredictable.
C. The temperature increases with the amount of long-wave radiation in our atmosphere.∗
D. Deforestation lowers the Earth’s average temperature.

Food Allergies

Diseases like pneumonia, tuberculosis, and meningitis are caused by bacteria, which have killed many people over the millions of years that humans have inhabited the Earth. Thus, it is no wonder that people generally associate bad things with bacteria. Yet, some bacteria that naturally exist on our skin, in the lining of our intestinal tract, and in our mucous membranes actually help protect us from diseases.

Bacteria are very simple organisms that are found throughout the environment, including in other living organisms, as well as in our foods that come from animals like cheese, meat, and milk. They are spherical or rod-shaped cells that generally measure several micrometers in length. Some bacterial cells are surrounded by the cell wall, which is a tough coat designed to protect the bacterium. Inside the cell wall is a cell membrane, which commonly encloses a single compartment where the DNA, proteins, and other small molecules are held. The protein in the bacteria produces enzymes, which are secreted and act on the organisms around them.
Antibiotics are drugs used to kill or harm specific bacteria. Since their discovery in the 1930s, antibiotics have made it possible to cure many diseases and save the lives of millions of people around the world. Many antibiotics work by inactivating an essential bacterial protein. Bacteria are always changing in an effort to resist the drugs that can kill them. Genetic change can alter the protein in the bacteria. Mutations in the target protein can prevent the antibiotic from binding to the protein, or if it does bind, prevent it from inactivating the target protein. Genetic change can also lead to increased production of the bacteria’s target enzymes so that there are too many of them and the antibiotic cannot inactivate them all.

When strains of bacteria become resistant to common antibiotics they are called “Superbugs.” Researchers are constantly creating new varieties of antibiotics to deal with the Superbugs. They are also creating new antibiotic treatments by combining previously used antibiotics. These new antibiotics are effective because they kill a broader range of bacteria and are called “broad spectrum” antibiotics.

It often comes as a surprise for us to learn that there are thousands of billions of bacteria living in the human digestive tract which weigh an amazing three and a half pounds. These bacteria represent many species. Different species live in different regions, which have varying degrees of acidity.

Bacteria in our digestive tract serve many beneficial functions. They manufacture B-vitamins and folic acid. They provide the enzyme lactase, which allows us to digest food. Through enzyme secretions, bacteria transform metabolic and microbial waste before they are discharged by the body. For example, the body creates bile not only as a lubricant to flush waste out of the liver, but also to detoxify many poisons accumulating in the liver. When bile enters the small intestine via the common bile duct, bacteria break down the bile salts into a less caustic compound. It becomes nondoingurous by the time it reaches the large intestine. Without the bacteria, the bile salts freely enter and damage the large intestine.

One of the most important functions of bacteria in our digestive tract is to keep in check the growth of pathogenic parasites, fungi, and yeasts such as Candida. In a healthy situation, the small intestine epithelium maintains tight cell junctions. This contributes to the physical barrier involved in intestinal absorption. In addition to the physical barrier, there is an important chemical barrier within the mucus. The mucus contains immune agents, which neutralize any toxin that comes in contact. The situation changes when there are not enough bacteria to control the growth of yeasts like Candida. Candida exudes an aldehyde secretion. These secretions make small intestine epithelial cells shrink, and the cell junctions to loosen. This allows intestinal toxins to infiltrate through the epithelium and pass into the blood. This prompts an immune response. The immune agents in the epithelial mucus remain the sole agent for neutralization. Eventually, the immune system becomes exhausted rising to this challenge.

When the integrity of the intestinal barrier has been compromised, intestinal toxins are not the only pathogens to be absorbed. The barrier, in a healthy state, selectively allows digested nutrients to enter the small intestine only once they are fully digested. When the barrier is compromised, nutrients can be absorbed before they are fully digested. Just like when intestinal toxins cross the barrier, this prompts an immune response. In this case, the body’s immune system will use specific antigen-antibody markers to tag these foods as foreign irritants. Once these foods are tagged as irritants, then an immune response will be mounted every time in the future that the particular food touches the epithelia. What started as a Candida irritation with shrinking of the epithelial cells has now been complicated with active inflammation every time a particular food is eaten. The allergic reaction to food particles can be manifested in skin rashes and inflammations.

1. Which of the following does bacteria NOT cause?
   A. Pneumonia
   B. Asthma
   C. Tuberculosis
   D. Meningitis

2. When were antibiotics discovered?
   A. 1890’s
   B. 1930’s
   C. 1950’s
   D. 1970’s

3. Candida is an example of a . . .
   A. bacterium
   B. parasite
   C. yeast
   D. fungus

4. How many bacteria do researchers estimate live in the human gastrointestinal tract?
   A. thousands
   B. hundreds of millions
   C. tens of billions
   D. thousands of billions

(Appendices continue)
5. What new pathogen is absorbed when the intestinal barrier is compromised by antibiotic use?
   A. bile
   B. undigested food particles
   C. folic acid
   D. lactase

6. When might damage to the small intestine epithelium occur?
   A. when epithelium cells grow in size
   B. when Candida multiplies in the small intestine
   C. when intestinal toxins are blocked from the bloodstream
   D. when food is absorbed before fully digested

7. When is a person most likely to develop food allergies?
   A. at the beginning of taking a course of antibiotics
   B. when bile enters the large intestine
   C. when the person has a high bacteria count in the digestive tract
   D. after completing a course of antibiotics

8. When undigested food particles cross the epithelium, when will a person experience an allergic reaction?
   A. immediately, as soon as the barrier is compromised
   B. only children will experience allergic reactions
   C. every time the food is eaten again
   D. only the first time the food is eaten

9. Samantha is taking a broad-spectrum antibiotic. Most likely her digestive tract will show an increase in
   A. vitamin B
   B. folic acid
   C. bacteria
   D. yeast

10. What might antibiotics that were developed to kill Superbugs lead to?
    A. skin rashes
    B. healthy small intestine epithelium
    C. blockage of intestinal toxins from the bloodstream
    D. absorption of nutrients after digestion occurs

IQ

(Note: This text was designed to argue for a particular viewpoint rather than reflect current scientific consensus.)

The idea that ethnic groups differ in intelligence lacks supporting evidence and is not logically plausible in light of certain facts. It is true that, on average, Asians score a couple of IQ points higher than Caucasians. In addition, Hispanics score several points lower than Caucasians, and African Americans score several points lower than Hispanics. However, these IQ scores do not provide evidence of real differences in any kind of general mental ability.

The first problem is that there is no such thing as a single “intelligence” that determines a person’s ability to think and make decisions in all situations. To behave intelligently means to adapt to one’s environment and situation and make appropriate choices. Different situations require different types of adaptations, and therefore different types of “intelligence.” Athletes, artists, doctors, and scientists all use different mental skills to perform their tasks well. People who can solve complex math problems are not the same people with the most common sense or people who solve social problems. IQ tests reduce many different aspects of thinking down to a single score. These many different abilities to perform very different tasks cannot be treated as a single ability.

Standard intelligence tests require the test-taker to interpret the words and concepts used in the problems and examples on these tests. Different ethnic groups, even within the U.S., experience very different environments, live in different parts of the country, and have different languages, vocabulary, or grammar. Intelligence tests were constructed and are generally scored by Caucasians. It is no surprise that many non-Caucasians do not perform as well on such tests.
Even the differences on limited and biased IQ tests are not caused by the biological differences between races. Ethnic groups differ genetically in only very superficial ways when a single gene could be involved, such as skin color, height, and other outward physical appearances. Solving complex mental problems involves many aspects of our thoughts, emotions, and behaviors that are the result of multitudes of genes interacting in complex ways. Also, even when two things are genetically identical they behave differently, because of their surrounding environment. For example, if you plant a handful of corn seeds in the fertile ground of Iowa and plant genetically identical seeds in the Arizona desert, the two groups of plants will come out very differently. Likewise, even identical twins differ in IQ, and siblings are often more different from each other than to people they are not related to.

The skills and knowledge that a child is exposed to in school play a large role in how the child performs on IQ tests. Quality of education depends upon circumstances, such as class size, teacher training, computers, and up-to-date books. In addition, parents and grandparents who obtained a better education will be better able to facilitate their child or grandchild’s education. The studies that show ethnic differences in IQ look at students who had already completed their primary education and were already affected by the quality of their education.

In addition to low-education quality, all school-related performance is harmed by home-life stress and hardships of poverty. Children who are hungry, malnourished, and concerned about family conflict will lack the focus and motivation needed to perform well on such tests. On average, most minority groups are lower in economic and social status compared with Whites.

The gap between Caucasians and minorities on intellectual tests such as SAT and ACT has been getting smaller over the past decades. Also, the average IQ score for all people today is several points higher than it was 50 years ago. Obviously, there is virtually no biological difference between children of today and their grandparents. Changes in scores over time are the result of changes in the economic status of minorities.

A more historical problem is that the American culture has lead minority groups to believe they are inferior and not as intelligent, for many generations. Low self-confidence will impact any kind of test that attempts to evaluate and compare people. Regardless of current economic success or education, most African Americans share this cultural history which subtly impacts their confidence in their own abilities.

1. How much higher do Asians score than Caucasians on IQ tests?
   A. 35
   B. 20
   C. 12
   D. 2

2. To behave intelligently means to . . .
   A. adapt to one’s environment and situation.
   B. make appropriate choices.
   C. make fewer mistakes than others.
   D. both A and B.

3. People who can solve complex math problems are not the same as people who can solve _________ problems.
   A. Scientific
   B. Social
   C. Engineering
   D. Physics

4. Which ethnic groups did the text specifically mention?
   A. Hispanics, Caucasians, African Americans, and Asians
   B. Asians, Caucasians, African Americans, and Arabs
   C. Asians, Caucasians, and African Americans
   D. Only Caucasians and African Americans

5. According to the text, which of the following does quality of education depend upon?
   A. class size and computers
   B. a safe, drug-free campus
   C. an ethnically diverse student body
   D. all of the above

6. Which of the following are true of scores on IQ tests?
   A. Asians score higher than Hispanics.
   B. Hispanics score higher than Caucasians.
   C. African-Americans score higher than Hispanics.
   D. All groups score the same.
7. Which of the following is implied by the text?

A. Grandparents will tend to have higher IQ scores than their grandchildren.
B. Grandparents will tend to have lower IQ scores than their grandchildren.
C. Grandparents will have the same IQ scores as their grandchildren.
D. IQ tests have only existed for 20 years, so no one knows.

8. Which of the following is NOT true, according to the text?

A. There is no such thing as a single general intelligence.
B. IQ tests are culturally biased.
C. Only African Americans who are poor will score lower than Caucasians.
D. None of the above, they are all true according to the text.

9. Which would be true if African Americans and Caucasians were compared in the first grade rather than in high school?

A. Their IQ scores would be more different.
B. Their IQ scores would be more similar.
C. African Americans would have higher IQs than Caucasians.
D. Both A and C.

10. What eliminates biological differences as a possible cause of differences between ethnic groups in solving complex problems?

A. African Americans have suffered historical injustices.
B. People who live in poverty receive a worse education.
C. IQ tests were created by Caucasians.
D. Ethnic groups only differ in ways that involve a single gene.

Monetary Policy

The U.S. is the largest economy in the world. Therefore, the U.S. dollar is considered a stable value. Many factors affect the strength of the dollar relative to other currencies, including the trade deficit or surplus, the size of the Government deficit, interest rates, and the strength of the U.S. economy. The strength of the dollar is also affected by the monetary policy imposed by the Federal Reserve System.

The Fed, as it is called, is the central bank of the United States. The Fed’s duties include conducting the nation’s monetary policy by influencing money and credit conditions in the economy in pursuit of full employment, stable prices, and promoting the stability of the financial system. The Fed conducts monetary policy using three major tools. It buys and sells U.S. treasury and federal agency securities in the open market; it sets the discount rate, which is the interest rate that banks pay the Fed to borrow money; and it sets reserve requirements, which is the amount of funds that banks must hold in reserve against deposits made by their customers. Monetary policy can affect short-term interest rates, foreign exchange rates, long-term interest rates, the amount of money and credit, and, ultimately, a range of economic variables, including employment, output, and prices of goods and services.

Monetary policy works by affecting the amount of money circulating in the economy. The Fed can change the amount of money that banks are holding in reserves by buying or selling existing U.S. Treasury Bonds. The Fed sells bonds, which decreases banks’ reserves and their ability to make loans. As banks lend less and the money supply decreases, interest rates increase. The Fed buys bonds, which increase banks’ reserves and their ability to make loans. As banks lend more and the money supply increases, interest rates decrease.

Lower interest rates mean that consumers pay less when they charge purchases. They may be more willing to spend. They may even buy expensive goods, like cars and refrigerators, to take advantage of lower interest rates. As the demand for more goods increases, either businesses will increase production to satisfy the demand or prices of goods will increase.

Lower interest rates may encourage businesses to expand to meet the increasing consumer demand. They may run extra shifts or build new factories. This may create new jobs. As workers, who were previously unemployed, return to the workforce, they will eventually spend their paychecks. This too will increase the demand for goods. Again, either businesses will increase production or prices of goods will increase.

(Appendices continue)
Sometimes consumer spending is so great that production can’t keep up with demand. The excessive demand for goods can lead to inflation. Inflation can also occur as a result of increasing the amount of money circulating in the economy. Inflation means dollars are worth less. The Fed will try to keep inflation in check.

Inflation may undermine the strength of the economy. Inflation increases the difficulty of forecasting prices and costs of doing business, so it discourages businesses from planning and investing. People also may be uncertain and reluctant to spend. Both of these factors could reduce the long-term level of economic growth. Inflation also increases the cost of carrying out transactions. Inflation in U.S. increases cost of U.S. goods; therefore, imports increase and exports decrease.

1. Which country has the world’s largest economy?
   A. China
   B. United Arab Emirates
   C. Japan
   D. United States

2. What is the Fed?
   A. the central bank of the U.S.*
   B. the Department of the Treasury
   C. the Department of Commerce
   D. the Securities and Exchange Commission

3. Which of the following does monetary policy affect?
   A. the amount of tariffs on foreign goods
   B. the amount of unemployment compensation available to citizens
   C. the amount of money circulating in the economy*
   D. the amount of money printed by the U.S. Treasury

4. What does inflation in the U.S. tend to result in?
   A. a decrease in imports of foreign goods
   B. a decrease in imports of foreign goods
   C. an increase in consumer spending
   D. an increase in the stability of the U.S. dollar

5. Which of the following is a cause of inflation?
   A. when long-term interest rates rise above short-term interest rates
   B. when production can’t keep up with consumer demand* 
   C. when production costs rise faster than the demand for goods
   D. when the price of stocks rises faster than earnings

6. If interest rates are lowered, consumers are more likely to
   A. buy more cars*
   B. buy more food
   C. save more
   D. travel less

7. Which is most likely to occur when the Fed increases the reserve requirement?
   A. Consumer spending will increase.
   B. Interest rates will increase.*
   C. Local banks will lend out more money.
   D. The economy will grow.

8. Unemployment will tend to decrease when . . .
   A. interest rates decrease*
   B. consumer demand decreases
   C. business investment decreases
   D. the money supply decreases

(Appendices continue)
9. Which of the following is NOT a likely result of lower interest rates?
   A. Prices of goods will decrease.
   B. Consumers are willing to spend more.
   C. Consumers will buy more expensive goods.
   D. Businesses will decrease production.

10. What might the Fed do if it wants to affect the economy in a way that is similar to that of lowering income taxes?
   A. Decrease loans to consumers and businesses.
   B. Decrease the reserve requirement.
   C. Increase the discount rate.
   D. Decrease the money supply.

Appendix B

**Average Monitoring Accuracy (with SDs) Analyzed as Gamma Correlations**

<table>
<thead>
<tr>
<th>Experiment and Condition</th>
<th>Metamemory</th>
<th>Metacomprension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1: Expectancy condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No expectancy</td>
<td>.35 (.09)</td>
<td>.16 (.09)</td>
</tr>
<tr>
<td>Memory</td>
<td>.52 (.06)</td>
<td>.07 (.08)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>.35 (.08)</td>
<td>.53 (.06)</td>
</tr>
<tr>
<td>Experiment 2: Expectancy—Manipulation type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory—Example tests</td>
<td>.57 (.09)</td>
<td>.08 (.12)</td>
</tr>
<tr>
<td>Comprehension—Example tests</td>
<td>.41 (.11)</td>
<td>.12 (.14)</td>
</tr>
<tr>
<td>Memory—Test description</td>
<td>.44 (.08)</td>
<td>.14 (.09)</td>
</tr>
<tr>
<td>Comprehension—Test description</td>
<td>.16 (.13)</td>
<td>.52 (.10)</td>
</tr>
<tr>
<td>Experiment 3: Postreading expectancy condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No expectancy</td>
<td>.36 (.09)</td>
<td>.08 (.08)</td>
</tr>
<tr>
<td>Memory</td>
<td>.39 (.10)</td>
<td>.27 (.12)</td>
</tr>
<tr>
<td>Comprehension</td>
<td>.13 (.13)</td>
<td>.46 (.11)</td>
</tr>
<tr>
<td>Experiment 4: Expectancy—Explanation condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No expectancy—No explanation</td>
<td>.34 (.07)</td>
<td>.08 (.08)</td>
</tr>
<tr>
<td>Comprehension—No explanation</td>
<td>.31 (.09)</td>
<td>.51 (.06)</td>
</tr>
<tr>
<td>No expectancy—Explanation</td>
<td>.19 (.08)</td>
<td>.46 (.05)</td>
</tr>
<tr>
<td>Comprehension—Explanation</td>
<td>.39 (.08)</td>
<td>.70 (.05)</td>
</tr>
</tbody>
</table>