8-1-2012

Elementary School Experience with Comprehension Testing May Influence Metacomprehension Accuracy Among Seventh and Eighth Graders

Keith W. Thiede
Boise State University

Joshua S. Redford
Boise State University

Jennifer Wiley
University of Illinois at Chicago

Thomas D. Griffin
University of Illinois at Chicago

Elementary School Experience with Comprehension Testing May Influence Metacomprehension Accuracy Among 7th and 8th Graders

Keith W. Thiede & Joshua S. Redford
Boise State University

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

Abstract

We explored whether exposure to different kinds of comprehension tests during elementary years influenced metacomprehension accuracy among 7th and 8th graders. This research was conducted in a kindergarten through eighth grade charter school with an expeditionary learning curriculum. In literacy instruction, teachers emphasize reading for meaning and inference building, and they regularly assess deep comprehension with summarization, discussion, dialogic reasoning and prediction activities throughout the elementary years. The school recently expanded, doubling enrollments in 7th and 8th grades. Thus, approximately half of the students had long-term exposure to the curriculum and the other half did not. In Study 1, metacomprehension accuracy using the standard relative accuracy paradigm was significantly better for long-time students than for newcomers. In Study 2, all students engaged in delayed-keyword generation before judging their comprehension of texts. Metacomprehension accuracy was again significantly better for long-time students than for newcomers. Further, the superior monitoring accuracy led to more effective regulation of study, as seen in better decisions about which texts to restudy, that led, in turn, better comprehension. The results suggest the importance of early exposure to comprehension tests for developing skills in comprehension monitoring and self-regulated learning.

Models of self-regulated learning describe learning as an interplay between metacognitive monitoring and regulation of study (e.g., Ariel, Dunlosky & Bailey, 2009; Butler & Winne, 1995; Griffin, Wiley & Salas, in press; Metcalfe, 2002; Nelson & Narens, 1990; Thiede & Dunlosky, 1999; Winne & Hadwin, 1998). For instance, as a person studies, he or she monitors his or her learning and uses this to guide subsequent study. If monitoring indicates that material has been adequately learned, he or she will stop studying. If monitoring indicates the material has not been adequately learned, he or she will continue to study. Thus, accurate monitoring is crucial for effective regulation of study (Winne & Perry, 2000). If a person does not accurately differentiate well-learned material from less-learned material, he or she could waste time studying material that is already well-learned or, even worse, fail to restudy material that has not yet been adequately understood.

Empirical support for the aforementioned models has largely been correlational. Metacognitive monitoring has been shown to be related to regulation of study. For example, the selections of items for restudy has been shown to be related to judgments of learning (e.g., Dunlosky & Thiede, 2004; Mazzoni & Cornoldi, 1993; Mazzoni, Cornoldi, & Marchitelli, 1990; Metcalfe, 2009; Nelson & Leonesio, 1988; Thiede & Anderson, 2003). Regulation of study has also been shown to be related to test performance. That is, decisions about which items to restudy influence subsequent test performance, both when regulatory decisions are made by the experimenter (e.g., Atkinson, 1972; Nelson, Dunlosky, Graf, & Narens, 1994) and when they are made by participants (Thiede, 1999).

To date, Thiede, Anderson, and Therriault (2003) have reported the only experimental study showing that better monitoring accuracy produces more effective regulation of study, which, in turn, leads to increased learning. In this study, college students engaged in the standard relative accuracy paradigm (Maki, 1998) in which they read a set of texts and then judged their comprehension of each text. Further, prior to judging their comprehension some students generated a list of five keywords that captured the essence of the text. Some generated keywords immediately after reading (the immediate-keyword group), some generated keywords after a delay from reading (the delayed-keyword group), and others did not generate keywords (the no-keyword group). After making judgments, all students took tests on each of the topics. Then they had the opportunity to select texts for restudy, and took a final set of tests.

These experimental conditions produced significantly different levels of relative monitoring accuracy (computed as the intra-individual correlation between each student's judgments of comprehension and their actual test performance). The delayed-keyword group was more accurate than the other two groups. Similarly, the differences in monitoring accuracy produced differences in regulation of study. As the delayed-keyword group more accurately distinguished less-learned texts from better-learned texts, they more effectively regulated their study (choosing to restudy texts that were less learned). By contrast, the other groups less accurately distinguished less-learned texts from better-learned texts, and less effectively regulated their study (essentially randomly selecting texts for restudy). Finally, the differences in regulation of study produced differences in learning. Overall comprehension as measured by performance on the final tests was greater for the delayed-keyword group than for the other groups. Thus, monitoring accuracy was shown to influence the effectiveness of regulation of study and subsequent learning. Given the importance of monitoring accuracy in learning, it is not surprising that a great deal of research has been dedicated to discovering ways to improve monitoring accuracy.

One context that has been receiving increasing attention is improving monitoring accuracy when students are learning from text. The term for this kind of monitoring accuracy is *metacomprehension accuracy*, and the theoretical underpinnings of the approaches used to improve metacomprehension accuracy combine models of metacognitive monitoring and comprehension (Rawson, Dunlosky, & Thiede, 2000; Weaver, 1990; Wiley, Griffin, & Thiede, 2005). In particular, the cue-utilization model of metacognitive monitoring (Koriat, 1997) and the construction-integration model of comprehension (Kintsch, 1998) provide a framework to understand which techniques might theoretically improve metacomprehension accuracy. Consider the processes involved in judging one's comprehension of texts. After reading, a person is asked to judge his or her comprehension of a text. According to the cue-utilization framework, the metacomprehension judgment may be based on a number of cues, such as how easily the text was processed during reading (Dunlosky & Rawson, 2005; Rawson & Dunlosky, 2002), how successfully the material had been retrieved at the time of the judgment (Baker & Dunlosky, 2006; Benjamin, Bjork, & Schwartz, 1998; Morris, 1990), the familiarity with the domain of the text (Glenberg & Epstein, 1987; Glenberg, Sanocki, Epstein & Morris, 1987; Griffin, Jee & Wiley, 2009; Maki & Serra, 1992), or global characteristics of texts such as length or difficulty (Weaver & Bryant, 1995). Metacomprehension accuracy will tend to increase as the cues that are used as a basis for comprehension judgments more highly correlate with performance on a test of comprehension (for empirical evidence linking metacomprehension accuracy and judgment cue basis, see Thiede, Griffin, Wiley, & Anderson, 2010).

The construction-integration model (Kintsch, 1988) suggests different cues may be available as a basis for metacomprehension judgments. According to this model, a reader creates multiple representations of a text as he or she reads. For instance, the reader constructs a representation of the surface level (i.e., the exact words), a textbase level (i.e., the meaning of sentences), and the situation-model level (i.e., connections between ideas contained in the text, and the connection between these ideas and prior knowledge). A well-constructed situation model integrates the ideas contained in a text and allows the reader to form a causal model including inferences implied by the text and predicted by the text. When tests of comprehension assess the quality of the situation model of a text (McNamara, Kintsch, Songer, & Kintsch, 1996), metacomprehension accuracy should increase if readers use cues based on their situation model to judge their comprehension (for a detailed discussion of designing texts and tests for assessing the situation model see Wiley et al., 2005).

Many of the techniques that have now been empirically shown to improve metacomprehension accuracy arguably focus readers on their situation model while judging comprehension. Consider the delayed-keyword effect (Thiede et al., 2003). It has been hypothesized that generating keywords after a delay provides cues that are predictive of performance on a test of comprehension. That is, in contrast to keywords generated immediately after reading, which could provide cues related to the surface features of a text, keywords generated after a delay, when memory for detail has faded from working memory, are more likely to provide cues related to the situation model of a text. Thus, the cues available for judging comprehension are more likely to be related to what will be tested. Consistent with this hypothesis, delayed keyword generation tasks have been shown to improve metacomprehension accuracy (Thiede et al., 2003; Thiede, Dunlosky, Griffin & Wiley, 2005).

Another approach that has been taken toward improving metacomprehension accuracy has been to align judgments and tests by manipulating encoding activities (e.g., Griffin, Wiley, & Thiede, 2008; Thomas & McDaniel, 2007). For instance, Thiede, Griffin, Wiley, and Anderson (2010) had college students construct concept maps, which focused participants on the connections of ideas among the texts while reading. Constructing concept maps made cues related to the construction of a situation model more available which again helped to align metacomprehension judgments with later comprehension tests. Again, aligning the basis for comprehension judgments with the demands of the upcoming comprehension tests improved metacomprehension accuracy.

More recently, Thiede, Griffin and Wiley (2011) improved metacomprehension accuracy by instilling comprehension test expectancies in college students prior to reading. In particular, students read a series of practice texts and took either memory tests (which assessed one's ability to remember details contained in a text) or inference tests (which assessed one's ability to connect ideas in a text, make conclusions, or generate predictions). They then read a new set of texts and judged their learning of the texts. Finally, students took both memory for details and inference tests. Monitoring accuracy was influenced by the test expectancy manipulation: For students expecting inference tests, judgments more strongly correlated with inference test performance than with memory test performance. By contrast, for students expecting memory tests, judgments more strongly correlated with memory test performance than with inference test performance.

Although the empirical results reported above are showing that several interventions have proven quite successful at improving metacomprehension accuracy among college-age samples (see Thiede, Griffin, Wiley & Redford, 2009 for a more complete review), much less is known about whether and how metacomprehension accuracy can be improved in younger readers. It is clear that younger readers are less skilled at judging their own understanding. While in college age populations, standard levels of metacomprehension accuracy are found to be around .27 (and manipulations can improve correlations to levels over .6 in the studies reported above), seventh graders appear to be substantially poorer at this skill. In particular, a recent study by Redford et al. (2011) found average relative metacomprehension accuracy among uninstructed seventh grade students was -.41 in one experiment and -.25 in the other (both were significantly different from zero). The uninstructed seventh graders in another recent study also had negative accuracy and were worse than random chance at predicting their relative performance (de Bruin et al., 2011). These negative correlations are showing that the seventh grade samples actually performed better on tests for texts they thought they had not understood, and worse on tests they thought they had understood.

In addition to poorer levels of accuracy in uninstructed students, it also appears that manipulations are less robust among younger students. Redford, Thiede, Wiley and Griffin (2011) recently attempted to improve metacomprehension accuracy by instructing 7th grade students to construct concept maps during reading of expository texts, but improvements were inconsistent across experiments. Although it appears that techniques that may improve monitoring accuracy for college students can also sometimes support better monitoring accuracy for 7th graders (de Bruin et al., 2011; Redford et al., 2011), improvements at this grade level are inconsistent. In addition to differences in levels of monitoring accuracy even with interventions in place, younger readers also do not appear to consistently regulate their study. That is, de Bruin et al., found that decisions about which texts to reread were related to metacomprehension judgments; whereas, Redford et al. found that decisions were not consistently related to metacomprehension judgments. Thus, it is not clear that 7th graders will use monitoring to guide regulation of study, and any improvements in monitoring accuracy that are realized may not translate into differences in regulation of study or subsequent learning. One reason why children may struggle to monitor their own learning from text, is that this kind of learning may demand more cognitive resources than simpler memorization tasks, which may leave fewer resources for monitoring (Griffin, Wiley & Thiede, 2009; Rawson, Dunlosky, & Thiede, 2000). This could be particularly important with children, as Roebers, von der Linden, and Howie (2007) showed that cognitive resources play an important role in children's monitoring. Due to this, it is an open question whether any conditions can be consistently shown to improve both monitoring accuracy and regulation of study among students of this age level.

An alternative proposal is that younger students' impoverished skills at predicting their own performance on upcoming comprehension tests may be more of a reflection of their testing experiences. Given that even college-age students tend to expect test items that focus on memory for details (Thiede, Wiley, Griffin & Anderson, 2010), we hypothesized that 7th grade students may also generally expect questions about reading assignments to assess

memory of details rather than inferences that could be made from the text. We conducted a pilot study to test this hypothesis in a typical, public school setting. We had students read a series of four texts and predict their performance on a five-item test (the nature of the test was not described to students). We then had students complete both inference tests and memory for detail tests (with the order of tests counterbalanced across students). We found that predictions were positively related to performance on the memory test (mean correlation = .40) and, as in previous studies, negatively related to performance on the inference tests (-.37). This suggests that 7th grade students in a typical public school curriculum do seem to have the expectancy that comprehension tests will ask them for memory for details rather than about inferences that can be drawn from the text.

In the present investigation, we took advantage of a naturally occurring situation to evaluate whether metacomprehension accuracy seems to be influenced by the experience students have with testing during their elementary school years. That is, this research was conducted in a charter school that uses a non-standard expeditionary learning curriculum (Campbell, Cousins, Farrell, Kamii, Lam, Rugen, & Udall, 1996). In literacy instruction at all grades, the curriculum emphasizes reading for meaning and inference building. Tests at all grades include assessments of deep comprehension, requiring students to generate inferences, conclusions, connections and predictions from the texts they read (e.g., writing summaries, constructing concept maps, engaging in Socratic discussions).

The school recently expanded enrollments, doubling enrollments in the 7th and 8th grades. Approximately half of the students had regular long-term exposure to tests of deep comprehension and the other half had long-term exposure to the more typical tests. As part of another study (Snow, Hoetker, Bremner, Oswalt, & Thiede, in preparation), we interviewed teachers at the charter school and another public school that serves a similar student population and is representative of the feeder schools from which the expanded enrollments came. To ascertain what key differences might exist between the enactment of the expeditionary curriculum and more traditional curricula, we surveyed teachers about their testing practices and also reviewed their classroom materials and assessments. We found that teachers at the charter school made a more concerted effort to evaluate deep comprehension at all ages. In elementary years, teachers at both schools focused instruction and testing on fundamentals for reading (e.g., phonics, decoding, and fluency); however, teachers at the charter school also reported spending time on metacognitive strategies for reading and seeing reading as meaning making. At the charter school, comprehension was often assessed by having students (a) write summaries of the materials that had been read, (b) predict what would happen next in a story (inference building), (c) discuss what had been read with the teacher (Socratic dialogues) or classmates (often as part of literature circles), and (d) discuss how the materials that had been read during reading instruction connected with ideas from other areas, such as social studies and science (these connections are emphasized throughout the expeditionary learning curriculum). By contrast, although teachers at the other school reported assessing deep comprehension (e.g., having students write summaries of materials they had read), they also reported assessing fundamental skills (e.g., fluency) that were the focus of district-level benchmarks for schools.

Moreover, curricular materials such as assessments were also examined. For teachers in the public school, most of the comprehension assessments were multiple-choice items that could not be answered without having read the material with attention to details (e.g., What color were Harry Potter's eyes?); whereas, in the charter school, a more typical comprehension assessment focused on the big picture -- students were often asked to write a summary of a text, which was to be shared and discussed with fellow students and the teacher.

If comprehension monitoring is affected by test expectancy, and test expectancy is affected by the tests to which students are exposed, then long-time students at the charter school should expect tests of deep comprehension requiring inferences, reasoning, and connections; whereas, newcomers to the school should expect tests of their memory of details. As a result, we predict that metacomprehension judgments will be more strongly related to inference test performance for long-time students than for newcomers. By contrast, we predict that metacomprehension judgments will be more strongly related to memory test performance for newcomers than for long-time students.

Because these studies take advantage of a naturally occurring context where only some students had long-standing exposure to a particular curriculum, students could not be randomly assigned to conditions as they would be in a true experiment. This of course raises concerns that the samples may not be matched on critical variables such as

reading ability or motivation for reading. Several measures are reported below, including teacher ratings of ability and initial test performance measures, which suggest that these samples were similar on these dimensions. In addition, it is very important to note that differences between the two samples on factors such as these could not actually impact the main dependent measure used in these studies, relative monitoring accuracy. These factors would be expected to impact overall test performance levels and possibly overall magnitude of confidence judgments. Two commonly used measures of judgment accuracy, absolute accuracy and confidence bias, are heavily dependent upon these overall magnitudes of performance and judgments (see Yates, 1990) and could indeed be biased by these differences. However, relative accuracy is a measure that is statistically independent of both overall test performance and judgment magnitude. Relative accuracy is a within-person correlation where accuracy depends upon selectively increasing or decreasing specific judgments for each tests to better match the pattern of relative test performance from test to test. Except in cases where overall magnitude produces a restricted range, relative accuracy is not open to direct influence by the various individual difference factors that would be expected to influence average performance and confidence levels (Nelson, 1984; Griffin, Wiley, & Salas, in press). Thus, the focus of this investigation is precisely on differences between samples on relative accuracy measures, and the main question is whether long-time exposure to a curriculum that routinely includes testing for deeper levels of understanding may lead to improved metacomprehension accuracy in seventh and eighth grade students.

Study 1

This study was designed to evaluate whether exposure to different kinds of testing affects comprehension monitoring. This study was conducted early in the school year (October); therefore, newcomers to the school had had limited exposure to testing that focused on deep comprehension.

Method

Participants

Seventy-one students participated in this study. All participants were treated in accord with APA ethical standards. Participants had either been at the school for a minimum of four years (long-time students; N = 31) or had just began at the school (newcomers; N = 40). Of the 31 long-time students, 17 were female and 14 were male, 17 were 7^{th} graders (ages 12 – 13) and 14 were 8^{th} graders (ages 13 – 14). Of the 40 newcomers, 22 were female and 18 were male, 22 were 7^{th} graders and 18 were 8^{th} graders. Subsequent to the study, teachers rated students' overall reading ability into three categories (superior, average, below average) based on performance during the fall semester. Overall reading ability did not differ across groups, $\chi^2(2) = 1.3$, $p = .53$.

Materials

Four science-based expository texts were adapted from passages appearing in junior high school science textbooks. Texts were chosen to represent distinct topics, from which an underlying complex causal relation or process could be extracted, that also afforded the creation of five detail-related questions. Each text was approximately 430 words long. As suggested by Wiley, Griffin, and Thiede (2005) the texts were developed so that the causal connections among ideas in the texts were not stated and needed to be generated by the reader. The readability of the texts was grade appropriate with Flesch-Kincaid grade levels ranging from 7.1 to 7.5. For each text, we constructed two sets of test items, each containing five questions (the texts and test sets can be found in the appendix). One set was designed to assess memory of the details explicitly stated in the text. The second set of items required participants to generate inferences about the ideas presented in the text (i.e. draw conclusions, make connections, generate predictions). Some of these items were constructed by creating concept maps of a text and then writing questions that required readers connect ideas that spanned different parts of the text. Others required reasoning from the text to generate conclusions or predictions. We did not attempt to make any fine-grained discrimination between types of inference items. What all of these items had in common is that they required processing beyond simple memory for information stated directly in the text (see also Hinze & Wiley, 2011 and Redford, et al. 2011).

Procedure

In this study, we employed the standard relative accuracy paradigm used by most studies in the metacomprehension literature (see Glenberg & Epstein, 1985; Maki 1998; Thiede, Griffin, Wiley & Redford, 2009). Participants were instructed that they would be reading four texts, judging how well they understood each text, and then answering test questions for each text. They were given an opportunity to ask questions about the procedure.

Participants read the four texts. After reading the last text, participants judged their comprehension for each text. The prompt for the metacomprehension judgment was, "Please indicate how many of the five questions you think you will answer correctly for the text entitled TITLE OF TEXT." Participants entered 0 to 5 for each rating. They were given no information about what particular kind of test to expect. Participants then answered two sets of test questions. They either answered inference questions for all four texts first and then memory for details questions, or they answered memory for details questions for all four texts first and then inference questions.

The presentation order of the four topics was held constant across each phase of the study (i.e., reading, judging, and testing) within each participant. The order of topics was counterbalanced across participants using a Latin Square design. Test sets were blocked by type to control for contamination from answering certain kinds of items on subsequent test performance. Test set order was counterbalanced across participants. Preliminary analyses showed that order of tests was not significant nor did it interact with the other independent variables (Fs < 1).

Design

As all participants completed both inference and memory for details test sets; thus, this was a within-participants variable. Participants were either long-time students or newcomers. Therefore, we had a 2 (kind of test: inference versus detail) x 2 (group: long-time versus newcomer) mixed design.

Results

Metacomprehension judgments and test performance. As metacomprehension accuracy is the relation between metacomprehension judgments and test performance, we first report data on these variables. The median of both metacomprehension judgments and test performances across the four texts was computed for each participant. We analyzed these data using the intra-individual means as well as the intra-individual medians. As the results were the same, we reported the analyses of the medians because it is the recommended measure of central tendency for small sets of scores where extreme scores may have an undue influence on the mean (Gravetter & Wallnau, 1999). The mean of the medians was then computed across participants in each group, see Table 1. The mean magnitude of metacomprehension judgments did not differ across groups, $t(69) < 1.00$, $p > .10$. More important for measures of relative accuracy, the variability did not differ across groups, $t(69) = 1.18$, $p = .24$.

Test performance was analyzed in a 2 x 2 ANOVA. There was a main effect for kind of test, $F(1, 69) = 41.95$, $MSe = .39$, $p < .001$, eta squared = .38, as all students did better on the memory for detail tests. There was not a main effect for group, $F(1, 69) = 1.95$, $MSe = .40$, $p = .17$. The interaction was significant, $F(1, 69) = 4.48$, $MSe = .39$, $p = .04$, eta squared = .06. Tests of simple effects revealed that performance on inference tests was significantly greater for long-time students than for newcomers, $F(1, 69) = 9.36$, $MSe = .39$, $p = .04$, eta squared = .12; whereas, performance on the memory for detail tests did not differ across groups, $F(1, 69) < 1$. More important for measures of relative accuracy, the variability did not differ across groups, $t(69) = 1.61$, $p = 12$.

Monitoring accuracy. As suggested by Nelson (1984), relative monitoring accuracy was operationalized as a Goodman-Kruskal gamma correlation between judgments and test performance. Therefore, for each participant, we computed two gamma correlations between judgments and test performance (one for inference tests and one for memory for detail tests) across the four texts[1]. The mean gamma was then computed across participants in the respective groups for each kind of test (see Figure 1). A major benefit of this approach to computing accuracy is that, unlike measures that simply take a difference score between judgments and performance, relative accuracy is not dependent upon the factors that impact a students' overall test performance or their overall level of confidence.

Instead, relative accuracy requires that students align their judgments to predict their own particular pattern of variance in performance from test to test.

Eight participants in each group had indeterminate gammas due to invariance in their metacomprehension judgments. We also conducted these analyses using an intra-individual Pearson correlations rather than gamma correlations. As the results were the same, we will report only the analyses with gamma to be consistent with the metacomprehension literature and for reasons discussed by Nelson (1984).

Monitoring accuracy was analyzed in a 2 x 2 ANOVA. There was a marginal main effect for kind of test, $F(1, 53) = 2.93$, $MSe = .53$, $p = .09$, eta squared = .05. There was a main effect for group, $F(1, 53) = 4.05$, $MSe = .50$, $p = .05$, eta squared = .07. The interaction was also significant, $F(1, 53) = 11.70$, $MSe = .53$, $p = .001$, eta squared = .18. Tests of simple effects revealed that monitoring accuracy for inference tests was significantly greater for long-time students than for newcomers, $F(1, 53) = 18.90$, $MSe = .53$, $p < .001$, eta squared = .26; whereas, monitoring accuracy for the memory for detail tests did not differ across groups, $F(1, 53) < 1$.

Discussion

Differences in monitoring accuracy suggest that the groups may have different expectations for the upcoming comprehension tests. In particular, judgments were positively correlated with inference test performance for long-time students, suggesting that their judgments were likely based on an expectation that comprehension would be assessed by tests of deeper comprehension rather than memory for details. By contrast, judgments were negatively correlated with inference performance for newcomers and positively related with detail test performance, suggesting that judgments for newcomers were likely based on an expectation that comprehension would be assessed by tests of memory for details rather than inference tests.

Study 2

The findings from Study 1 indicate differences in metacomprehension accuracy between groups, which set the stage for evaluating the effect of monitoring accuracy on regulation and learning in these samples. As noted in the introduction, only one study (Thiede et al., 2003) has shown the effect of monitoring accuracy on regulation of study and subsequent learning. Study 2 attempts to extend the findings of Thiede et al. in that it examines the effect of monitoring accuracy on regulation and learning among 7th and 8th grade students (rather than college students).

This study was conducted on the same samples of charter school students during the spring semester (May) of the same academic year as the first study. Although the newcomers had been exposed to tests emphasizing deep comprehension throughout their first year at the charter school, their exposure was considerably less than that experienced by long-time students. Therefore, we maintained enrollment group as a between participants variable. The findings from Study 1 showed clearly that metacomprehension accuracy differed among these two groups of students. However, the levels of metacomprehension accuracy were not particularly high, which might affect whether students use comprehension monitoring to make decisions about which text to reread (i.e., they may discount the utility of comprehension monitoring if they are not confident their monitoring is accurate). Therefore, we employed a delayed-keyword generation task (see Thiede et al., 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005) to improve levels of metacomprehension accuracy for both groups.

If monitoring judgments more accurately reflect inference test performance for long-time students than for newcomers, we would expect to see more effective regulation of study among long-time students. Further, we would expect to see greater improvements in comprehension (i.e., better inference test performance on a final test) after restudy opportunities for the long-time students -- as a result of more effective regulation of study. Thus, the key dependent measures for this study include metacomprehension accuracy, but also regulation of study, and final learning outcomes.

Method

Participants

Seventy students participated in this study and were the same as in Study 1.

Materials

Four new texts and tests were created for Study 2. The texts were similar in construction to those used in Study 1 and were on the topics of breeding and cloning, energy from food, bacteria, and the carbon cycle.

Procedure

The procedure largely followed that of Study 1 with a few key changes. Participants were instructed that they would be reading four texts, judging how well they understood each text, and then answering test questions for each text. They were also instructed that they would be writing a list of five keywords that captured the essence of the text prior to judging their comprehension. These instructions included an example of keywords for a text: "For example, if you had read a text about the Titanic, you might generate the keywords: Iceberg, Shipwreck, Tragedy, and so on." Participants were also instructed that following the first set of tests, they would select one text for rereading, with the goal of "maximizing your overall test score across the four texts." Following the instructions, participants were given an opportunity to ask questions about the procedure.

In this study, after reading the four texts, participants generated five keywords for each text. Reading all texts before generating any keywords provided a delay between reading and generating for each text, which has been shown to be essential to improving metacomprehension accuracy (Thiede, Dunlosky, Griffin, & Wiley, 2005). After generating keywords for the last text, participants judged their comprehension for each text and then completed inference tests for each text. Participants then selected *one* text for restudy. We chose to have participants select a text for restudy after completing just the inference tests because we wanted to hold test expectancy constant across groups. That is, if test expectancy affects metacognitive monitoring (e.g., Thiede, et al, 2011) it may also effect regulation of study; therefore, failure to hold test expectancy constant could make the regulation data difficult to interpret. We chose to create the expectancy of inference tests in this study because we (and the teachers) wanted to examine regulation for deeper comprehension rather than memory of texts. After selecting a text for rereading, participants answered memory questions for each text before being given the text they had selected for rereading. This allowed us to examine accuracy of monitoring memory without contamination from rereading. Participants then reread the text that they had selected. Finally, participants answered inference questions and then memory questions for the text they reread.

Design

As in Study 1, we had a 2 (kind of test: inference versus detail) x 2 (group: long-time versus newcomer) mixed design.

Results

Metacomprehension judgments and test performance. As in Study 1, we first report data on metacomprehension judgments and initial test performance. Final test performance for the text selected for rereading will be reported below. The median of both variables was computed across the four texts for each participant. The mean of the medians was then computed across participants in each group, see Table 1. As in Study 1, results using the intra-individual means were the same; therefore, we report only the results based on the intra-individual median. The mean magnitude of metacomprehension judgments did not differ across groups, $t(68) < 1.00$, $p > .10$. More important for measures of relative accuracy, the variability did not differ across groups, $t(68) < 1.00$, $p > .10$.

Test performance was analyzed in a 2 x 2 ANOVA. There was a main effect for kind of test, $F(1, 68) = 40.88$, $MSe = .41$, $p < .001$, eta squared = .38. There was not a main effect for group, nor was the interaction significant, both $F(1, 68) < 1$. Regarding the main effect for kind of test, as seen in Table 1, test performance was greater for memory for detail tests than for inference tests. However, there was no difference between long-term students and newcomers in test performance. It is important to note the lack of difference in test performance across the two groups suggests that these samples do not differ in basic reading proficiency or motivation. More important for measures of relative accuracy, the variability did not differ across groups, $t(68) < 1.00$, $p > .10$.

Monitoring accuracy. For each participant, as in Study 1 we computed two gamma correlations between metacomprehension judgments and test performance. The mean gamma was then computed across participants in the respective groups for each kind of test (see Figure 2). Two participants in each group had indeterminate gammas due to invariance in their metacomprehension judgments. We also conducted these analyses using intra-individual Pearson correlations rather than gammas, as the results were the same, we report only the results using gamma.

Monitoring accuracy was analyzed in a 2 x 2 ANOVA. Neither main effect was significant, both $F(1, 64) < 1.3$. However, the interaction was significant, $F(1, 64) = 18.70$, $MSe = .49$, $p = .001$, eta squared = .23. Tests of simple effects revealed that monitoring accuracy for inference tests was significantly greater for long-time students than for newcomers, $F(1, 64) = 14.61$, $MSe = .50$, $p < .001$, eta squared = .19; whereas, monitoring accuracy for the memory for detail tests was significantly greater for newcomers than for long-time students, $F(1, 64) = 4.78$, $MSe = .50$, $p = .03$, eta squared = .07.

Regulation of study. As is common in the metacognitive literature, regulation of study was operationally defined as the correlation between metacognitive judgments and selection of an item for restudy (e.g., Nelson, Dunlosky, Graf, & Narens, 1994; Thiede & Dunlosky, 1999). Selection was coded as a 1 for the text selected for rereading and 0 for those not selected for rereading. Thus, a negative correlation indicates that a participant chose to reread a text that was judged to be less well understood. For each participant, we computed a gamma correlation between metacomprehension judgments and text selection. The mean gamma was then computed across participants in the respective groups. Two participants in each group had indeterminate gammas due to invariance in their metacomprehension judgments.

Regulation differed significantly across groups, $t(64) = 2.99$, $p = .004$. As seen in Figure 3, regulation was more strongly negative for the long-time students than for the newcomers. Thus it appears that superior monitoring accuracy led to more adaptive or appropriate regulation of study. Put differently, the long-time students knew what they did not understand and they compensated for this with additional study of a lesser known text. By contrast, the newcomers did not know what they understood and essentially selected texts randomly for rereading (as indicated by the regulation correlation near 0).

Another way to evaluate differences in regulation of study is to examine the mean test performance for texts that were selected for restudy versus those that were not selected for study. Adaptive regulation would attempt to compensate for poor initial comprehension by allocating additional study time to those texts (e.g., Nelson et al., 1994). Thus, initial test performance for texts selected for restudy should be less than initial test performance for texts not selected for restudy. With this in mind, we evaluated differences in regulation of study by comparing mean test performance for texts selected for restudy and those not selected for restudy across the two groups—see Table 2. In particular, we conducted a 2 (selection: selected versus not selected) x 2 (group: long-time students versus newcomers) ANOVA for initial inference test performance and another for initial memory for detail test performance.

For inference test performance, there was a main effect for selection, $F(1, 68) = 7.79$, $MSe = .91$, $p = .007$, eta squared = .10. There was not a main effect for group, $F(1, 68) = 1.69$, $MSe = 1.13$, $p = .20$. The interaction was also significant, $F(1, 68) = 5.68$, $MSe = .91$, $p = .02$, eta squared = .08. Tests of simple effects revealed that, for long-time students, initial test performance was significantly lower for texts selected for restudy than for texts not selected for restudy, $F(1, 29) = 11.65$, $MSe = .92$, $p = .002$, eta squared = .29. By contrast, for newcomers, initial test performance did not differ across texts that were selected for restudy and those not selected for restudy, $F(1, 39)$

< 1. These data suggest that long-time students (who were more accurately monitoring their inference performance) more appropriately regulated their study than did newcomers.

We also examined regulation based on memory performance. The 2 (selection: selected versus not selected) x 2 (group: long-time students versus newcomers) ANOVA revealed a main effect for selection, $F(1, 68) = 16.95$, $MSe = 1.05$, $p < .001$, eta squared = .20. Neither the main effect for group nor the interaction were significant, both Fs < 1. As seen in the bottom section of Table 2, mean initial test performance was worse for texts selected for restudy than for texts not selected for restudy for both groups. These findings suggest that both groups appropriately regulated study based on memory performance.

Final test performance on the text selected for rereading. Differences in regulation are hypothesized to affect learning, with more adaptive regulation leading to superior learning (Thiede et al., 2003). A 2 (group: long-time students versus newcomers) x 2 (pretest versus posttest) x 2 (kind of test: inference versus memory for detail) revealed a three-way interaction, $F(1, 68) = 24.06$, $MSe = .30$, $p < .001$, eta squared = .26. To better understand the three-way interaction, we conducted a 2 (group: long-time students versus newcomers) x 2 (pretest versus posttest) ANOVA for each kind of test separately.

For inference tests, there was significant interaction, $F(1, 68) = 25.62$, $MSe = .58$, $p < .001$, eta squared = .27. Follow-up tests of simple effects revealed that inference test performance increased significantly from pretest to posttest for the long-time students, $F(1, 29) = 49.29$, $MSe = .57$, $p < .001$, eta squared = .63. However, inference test performance did not significantly change from pretest to posttest for newcomers, $F(1, 39) < 1$—see Table 3. These findings demonstrate that long-time students engaged in more effective self-regulated learning, as they made more adaptive restudy decisions that in turn led to better comprehension as measured by the final inference tests.

For memory for detail tests, there was a main effect for pretest/posttest, $F(1, 68) = 27.0$, $MSe = .11$, $p < .001$, eta squared = .28. Neither the main effect for group nor the interaction were significant, both $F(1, 68) < 1$. As seen in Table 2, memory for detail test performance increased equally for the groups from pretest to posttest.

Discussion

The results of Study 2 show that as in Study 1, the long-time students continued to have better metacomprehension accuracy than the newcomers. At the same time, the average judgments of comprehension and average test scores did not differ between these groups. As in Study 1, the strong positive correlation between metacomprehension judgments and inference test performance, but not memory test performance, for long-time students suggests that their judgments were likely based on an expectation that comprehension would be assessed by tests of deeper comprehension rather than memory for details. By contrast, the strong positive correlation between metacomprehension judgments and memory for detail test performance for newcomers suggests that their judgments were likely based on an expectation that comprehension would be assessed by tests of memory for detail rather than deeper comprehension.

Monitoring accuracy is important because this information informs study decisions (regulation of study), and the restudy decision data in Study 2 show the effect of monitoring accuracy on regulation behaviors. Perhaps more important, the difference in regulation behaviors also affected comprehension as a result of restudy. The long-time students more accurately monitored their comprehension and more effectively regulated their study than did the newcomers—and this produced superior test performance on final inference tests, with no detriment to memory performance.

General Discussion

Many models of self-regulated learning suggest that metacognitive monitoring plays a key role in learning, in that it provides information to guide regulation of study, which in turn affects learning (e.g., Griffin, Wiley & Salas, in press; Thiede & Dunlosky, 1999; Winne & Hadwin, 1998). Although correlational data have suggested a relation between monitoring, regulation of study and learning, before the present study, only one study had shown experimentally the importance of monitoring accuracy for subsequent learning (Thiede, Anderson, & Therriault,

2003). Thiede, Anderson, and Therriault (2003) produced differences in monitoring accuracy among college students, using different experimental conditions: a no-keyword condition, an immediate-keyword condition, and a delayed-keyword condition—which had greater monitoring accuracy than the other conditions. Differences in monitoring accuracy led to more effective regulation of study (better decisions about which texts to reread), which in turn led to better overall reading comprehension.

The results of Study 2 are important because they replicate the findings of Thiede et al. (2003) and extend them to 7th and 8th grade students. Although previous research was not clear regarding whether 7th or 8th graders could engage in accurate monitoring or regulation, these data show a strong relation between restudy decisions and metacomprehension judgments among the long-time students. Further, levels of metacomprehension accuracy achieved in these students were more than twice that of the best accuracy levels reported in previous research with this age group; in de Bruin et al. (2011) the highest accuracy was .27, and in Redford et al. (2011) the best accuracy was .34. Study 2 demonstrates that even younger students will learn more if they can more accurately monitor their learning during study and use this information to regulate their learning. The results of Study 2 also suggest that if younger students do not accurately monitor their learning, as was the case with newcomers to the school, they will less effectively regulate their learning and fail to increase their comprehension of texts through additional study. Thus, it is crucial that we find ways to improve metacomprehension accuracy.

The cue-utilization framework of metacognitive monitoring (Koriat, 1997) suggests that monitoring accuracy improves as the cues used to judge one's learning are predictive of test performance. In these studies, we did not ask long-time students or newcomers to describe the bases of their judgments of comprehension; therefore, we have no direct evidence that the groups made use of different cues in judging comprehension. Moreover, if the groups did have access to different cues, we cannot be certain what produced the different cues. That is, although we have attributed differences in monitoring accuracy between the two groups to expectations created via different testing experiences (deeper comprehension versus surface memory for text), it may be that differences are attributable to other factors. Perhaps the increased emphasis on writing or production of graphic representations of texts in the charter school curriculum produced cues that were more predictive of performance on a test of deep comprehension. Perhaps something unrelated to reading curriculum (emphasis on different kinds of learning in the science curriculum) led long-time students to attend to different cues when judging their level of learning. Future research is needed to ascertain the source of cues used by younger readers (cf. Thiede et al., 2010, which obtained self-report data on cue bases for college students) before any direct conclusions can be drawn about this. That said, the specific differences seen *only* in monitoring accuracy for inference test performance suggests that long-time students and newcomers are judging their understanding of texts in different ways and against a different standard, and that newcomers' metacomprehension judgments are not well aligned to inference test performance. At a general level, various features of the expeditionary curriculum and its enactment may share the feature of emphasizing an expectation of what it means to understand a science text that goes beyond memory for factual details. It is this difference in expectations that provides a compelling account for the present pattern of results. As explained below, a number of other potential differences between the groups do not satisfy the various criteria for being able to account for relative accuracy differences specific to comprehension monitoring, especially in light of complete pattern of data across studies and measures.

In order to explain these differences in relative accuracy, any account would need to explain how the mechanism could lead the students to selectively alter certain judgments, in a particular pattern that better matched the way in which their test performance would vary on inference items. In addition, the assumed mechanism cannot be one that would be expected to also lead to differences in initial test performance or overall judgment magnitude, given that these variables did not differ between the groups of students. Yet another constraint is that any explanation must account for the fact that long-time students made judgments that better predicted inference performance but were less predictive of memory test performance. These various constraints mean that the majority of possible differences between the groups are not viable candidates for explaining the differences in relative metacomprehension accuracy, especially group differences that would be related to a students' overall achievement or confidence. In contrast, having prior experience with test formats that assess inferences and deeper comprehension is a viable candidate. This kind of knowledge can create expectations that allow students set a more appropriate standard against which to judge their likely performance on each particular topic. Such an influence would be not be expected to raise or lower judgments overall or to produce superior performance on tests after only a single reading (i.e. without an

opportunity to regulate study). They would however be expected to lead to judgments that better predict inference test performance but not memory test performance, and to selective restudy texts that they understood poorly, and to improve their performance on inference tests for those texts that they restudied. Thus, although these non-randomly assigned groups may differ in a number of potential ways, few if any alternative explanations seem to provide as coherent an argument for the full pattern of data across the measures than the differences in exposure to tests that emphasized deeper comprehension rather than memory.

The results of this investigation suggest that long-term exposure to tests that assess deeper comprehension can affect the way readers monitor comprehension. Or stating the inverse, failure to regularly assess deeper comprehension can lead readers to monitor memorization of texts, which has a detrimental effect on their later ability to engage in effective self-regulated learning—as shown by the differences in restudy decisions as well as final inference test performance in Study 2. Finding ways to infuse early reading curricula with opportunities to read-for-understanding would seem to be an important implication of this work.

References

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 133,* 432-447.

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology, 96,* 124-129.

Baker, J. M. C., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review, 13*, 60-65.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55 – 68.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65,* 245-281.

Campbell, M., Cousins, E., Farrell, G., Kamii, M., Lam, D., Rugen, L., & Udall, D. (1996). The expeditionary learning outward bound design. In S. Stringfield, S. M. Ross, & L. Smith (Eds.). *Bold Plans for School Restructuring: The New American Schools Designs.* (pp. 109-138). Mahwah, NJ: LEA.

de Bruin, A., Thiede, K. W., Camp, G., & Redford, J.R. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*, 294-310.

Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology, 47,* 274-296.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16,* 228-232.

Dunlosky, J., & Rawson, K. A. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes, 40*, 37-56.

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory & Cognition, 32*, 779-788.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11*, 702-718.

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

Goodman, L.A., & Kruskal, W. H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association, 49*, 732-764.

Gravetter, F. J., & Wallnau, L. B. (1999). *Essentials of statistics for the behavioral sciences* (3rd Ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

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

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

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36,* 93-103.

Hinze, S. R., & Wiley, J., (2011). Testing the limits of testing effects using completion tests. *Memory, 19,* 290-304.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95,* 163-182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kobasigawa, A., & Metcalf-Haggart, A. (1993). Spontaneous allocation of study time in first- and third-grade children in a simple memory task. *Journal of Genetic Psychology, 154,* 223-235.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370.

Maki, R. H. (1998). Test predictions over text material. In Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). *Metacognition in Educational Theory and Practice.* (pp. 117-144). Hillsdale, NJ: LEA.

Maki, R. H., & Serra, M. (1992). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18*, 116-126.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122,* 47-60.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18,* 196-204.

McDaniel, M. A., & Einstein, G. O. (1989). Material appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review, 1,* 113-145.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1–43.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131,* 349-363.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18,* 159-163.

Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16,* 223–232.

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

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5,* 207-213.

Nelson, T. O., & Leonesio, J. (1988). Allocation of self-paced study time and the "Labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14,* 676-686.

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, vol. 26, (pp. 125-173). New York: Academic Press.

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28,* 69-80.

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

Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction*, doi:10.1016/j.learninstruc.2011.10.007

Roebers, C. M., von der Linden, N., & Howie, P. (2007). Favourable and unfavourable conditions for children's confidence judgments. *British Journal of Developmental Psychology, 25*, 109-134.

Snow, J., Hoetker, G. A., Bremner, A., Oswalt, S., & Thiede, K. W. (in preparation). The influence of curriculum on classroom assessment.

Thiede, K.W. (1999). The importance of accurate monitoring and effective self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6*, 662-667.

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28,* 129-160.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95,* 66-73.

Thiede, K.W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 25*, 1024-1037.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *31*, 1267-1280.

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

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

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

Thomas, A. K. & McDaniel, M. A. (2007). The negative cascade of incongruent task-test processing in memory and metamemory. *Memory & Cognition, 35*, 668-678.

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

Weaver, C. A., III, & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition, 23*, 12-22.

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

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds). *Metacognition in Educational Theory and Practice.* (pp. 277-304). Hillsdale, NJ: LEA.

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts & P. Pintrich (Eds.), *Handbook of Self-Regulation, pp. 531-566.* New York: Academic Press, Inc.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall.

Author Notes

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through R305B07460 to Keith W. Thiede, Jennifer Wiley, and Thomas D. Griffin. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors thank Scott Hinze for his assistance in the creation of texts and tests.

Footnotes

1. Nelson (1984) recommended using a Goodman-Kruskal gamma correlation (Goodman & Kruskal, 1954) for these kinds of data. Gamma is computed by examining the direction of one variable relative to another. If one variable (e.g., metacomprehension judgment) is increasing from one text to another and the other variable (e.g., test performance) is also increasing across this same pair of texts, this is considered a concordance (C). By contrast, if one variable is increasing from one text to another and the other variable is decreasing across this same pair of texts, this is considered a discordance (D). Concordance and discordance is computed across all pairs of items. The total number of each is used to compute the correlation coefficient, Gamma = (C – D)/(C + D). Once computed, the Gamma coefficients then represent a continuous and normally distributed measure of judgment-performance correspondence suitable for analysis with most GLM approaches.

Table 1

Mean Metacognitive Judgments and Initial Test Performance by Group

Group	Judgment Magnitude	Detail Test Performance	Inference Test Performance
Study 1			
Long-time	3.66 (.14)	3.48 (.15)	3.01 (.09)
Newcomer	3.57 (.13)	3.55 (.10)	2.64 (.08)
Study 2			
Long-time	3.88 (.13)	3.57 (.17)	2.83 (.13)
Newcomer	3.91 (.11)	3.54 (.15)	2.88 (.11)

The entries are the mean metacognitive judgment and test performance computed across participants within each condition. The numbers in parentheses are the standard errors of the means.

Table 2

Mean Performance for Initial Inference and Memory Tests by Group and by Texts that were Selected versus Not Selected for Rereading

Group	Selected	Not Selected
Inference Test		
Long-time	2.20 (.21)	3.05 (.15)
Newcomer	2.83 (.18)	2.89 (.13)
Detail Test		
Long-time	3.00 (.16)	3.76 (.22)
Newcomer	3.03 (.13)	3.71 (.19)

The entries are the mean test performance computed across participants within each condition. The numbers in parentheses are the standard errors of the means.

Table 3

Mean Performance for Final Inference and Memory Tests by Group for the Text Selected for Rereading

Group	Pretest Performance	Posttest Performance
Inference Test		
Long-time	2.20 (.21)	3.57 (.17)
Newcomer	2.83 (.18)	2.88 (.15)
Detail Test		
Long-time	3.00 (.16)	3.30 (.16)
Newcomer	3.03 (.13)	3.30 (.14)

The entries are the mean test performance computed across participants within each condition. The numbers in parentheses are the standard errors of the means.

Figure Captions

Figure 1. Mean monitoring accuracy for each test by group in Study 1. The error bars represent the standard error of the mean.

Figure 2. Mean monitoring accuracy for each test by group in Study 2. The error bars represent the standard error of the mean.

Figure 3. Mean regulation of study by group in Study 2. The error bars represent the standard error of the mean.

Figure 1.

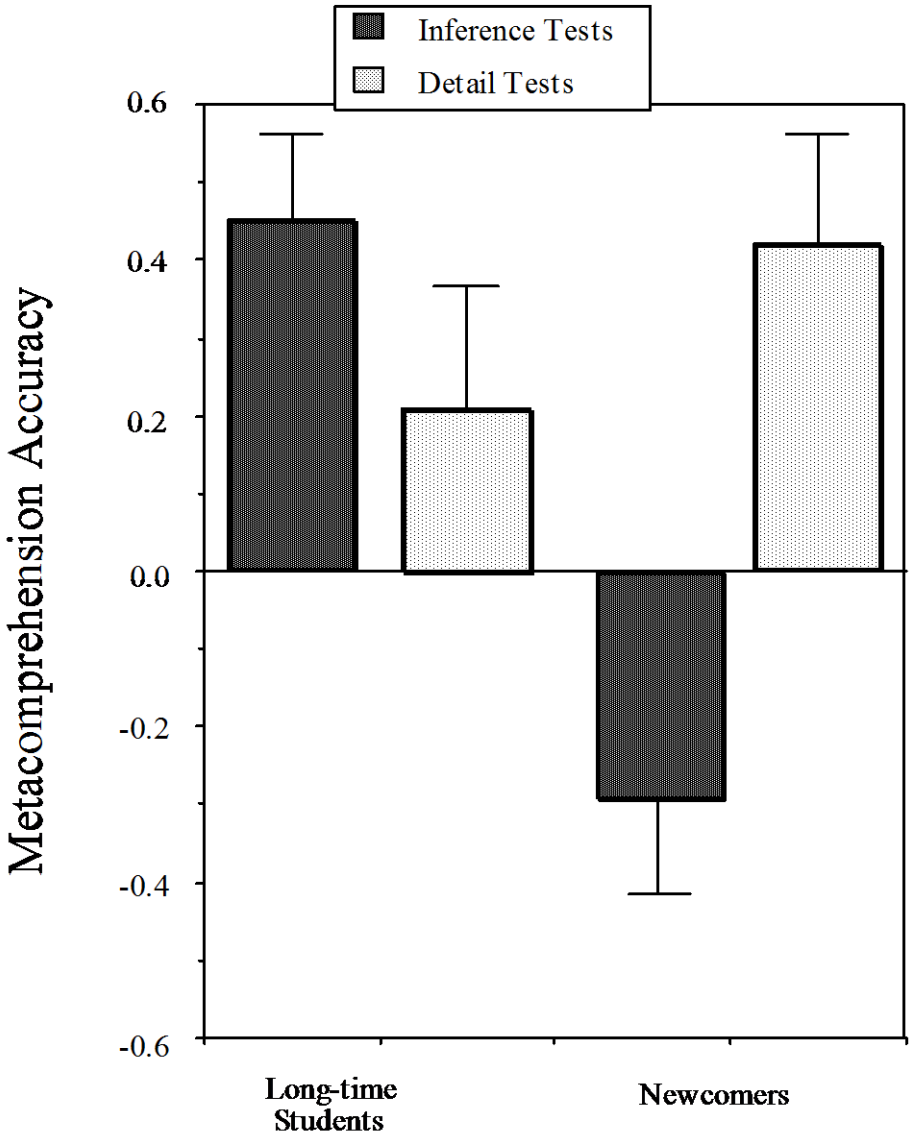

Figure 2.

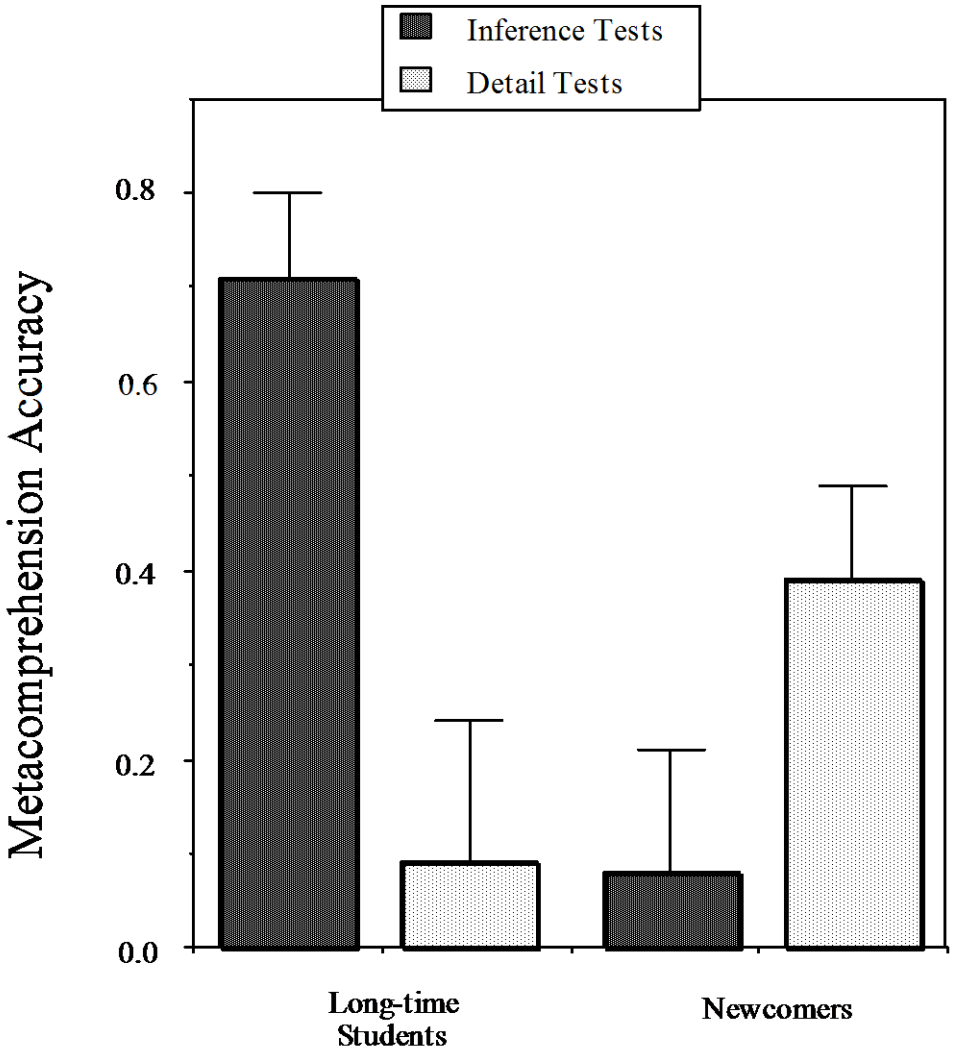

Figure 3.