

# Why Does Rereading Improve Metacomprehension Accuracy? Evaluating the Levels-of-Disruption Hypothesis for the Rereading Effect

John Dunlosky

*The University of North Carolina at Greensboro*

Katherine A. Rawson

*University of Colorado at Boulder*

Rereading can improve the accuracy of people's predictions of future test performance for text material. This research investigated this *rereading effect* by evaluating 2 predictions from the *levels-of-disruption* hypothesis: (a) The rereading effect will occur when the criterion test measures comprehension of the text, and (b) the rereading effect will not occur when a 1-week delay occurs between initial reading and rereading. Participants ( $N = 113$ ) were assigned to 1 of 3 groups: single reading, immediate rereading, or rereading after a 1-week delay. Outcomes were consistent with the 2 predictions stated earlier. This article discusses the status of the levels-of-disruption hypothesis and alternative hypotheses based on the cognitive effort required to process texts.

Metacomprehension involves a person's assessment of his or her own comprehension of text, which is a central component of self-regulated comprehension (Hacker, 1998). The role of metacomprehension in self-regulation is illustrated when a student is studying for an upcoming exam. After reading a textbook chapter, the student may assess how well he or she understands each section, which in turn provides input for further regulation. For example, if the student judges that a section is not well understood, he or she may study it longer (cf. Son & Metcalfe, 2000). Accordingly, the effectiveness of self-regulated comprehension is partly

determined by the accuracy of the metacomprehension judgments (Thiede, Anderson, & Theriault, 2003): If the judgments are not indicative of actual comprehension, the person may unnecessarily reread material that is relatively well understood and (even worse) fail to study material that is not well understood.

Therefore, how accurate are people at assessing their comprehension of text? About two decades of research on metacomprehension has consistently demonstrated that the accuracy of people's judgments of text comprehension is relatively low. More specifically, correlations between judgments and measures of text comprehension have usually been below  $+0.30$  (for reviews, see Maki, 1998; Weaver, Bryant, & Burns, 1995).

To provide insight into why metacomprehension accuracy may often be so poor, Rawson, Dunlosky, and Thiede (2000) integrated theory of metacognitive monitoring with theory of text comprehension. Their proposal integrated a cue-based account of metacognitive judgments with standard assumptions about the levels of text representation. According to this account, metacomprehension judgments are presumably based on a variety of cues such as domain familiarity, momentary accessibility, or processing ease (Maki, 1998). One influential cue is the relative amount of disruption in comprehension processes that occurs while an individual is reading a text, with more disruptions resulting in lower judgments of comprehension (Rawson & Dunlosky, 2002). For instance, an individual's text processing may be disrupted multiple times when reading some texts (e.g., when reading texts that are poorly written, or when the reader does not have the background knowledge needed to interpret text content) but rarely disrupted when reading others. After an individual reads a paragraph and is asked to make a metacomprehension judgment, he or she, in part, judges how well the paragraph was understood by estimating the frequency of disruptions that had occurred while reading, regardless of why those disruptions arose. Presumably, the individual would make lower comprehension judgments for texts resulting in more disruptions than for those resulting in fewer disruptions (Rawson & Dunlosky, 2002).

According to this cue-based account of metacognitive judgments (Koriat, 1997), judgment accuracy will be relatively high if the cues—in this case, disruptions—on which the judgments are based are highly correlated with measures of text comprehension. Therefore, a critical question arises: When will processing disruptions across texts be predictive (or diagnostic) of the relative comprehension performance for those texts? To provide an answer to this question, we turn to theory of text comprehension, which elucidates both the kinds of disruption that may occur during comprehension and when those disruptions will be predictive of comprehension.

Based on comprehension theory, a reasonable assumption is that the diagnosticity of processing disruptions for predicting test performance will depend on the level of text representation for which processing is disrupted. While reading a text, an individual can process the surface level representation (i.e., explicit linguis-

tic information), the text-base representation (i.e., semantic propositions and their interconnections), and the situation model (i.e., a representation of the situation described by the text). Disruptions may occur while an individual is processing at any of these levels.<sup>1</sup> For example, an individual's processing may be disrupted by difficult words, by attempting to resolve anaphors, or by generating a mental model of the situation described in the text. Most important, for now, is the processing involved in creating a situation model because performance on tests of comprehension is heavily dependent on the situation level of representation (Kintsch, 1994). Based on this rationale, the *levels-of-disruption* hypothesis asserts that judgments based on disruptions that occur when an individual is processing text at the situation level are expected to be relatively predictive of test performance across texts.

To better understand this expectation, consider an individual processing texts at the situation level. Disruptions in processing will likely occur less often for some texts than for others, and an individual's judgments of comprehension presumably will be lower for those texts in which a greater number of disruptions occurred. Assuming that these disruptions in processing the situation model also result in lower comprehension, performance on a test of comprehension will be lower for those texts in which many disruptions occurred. Therefore, when an individual is processing at the level of the situation model across texts, processing disruptions will not only influence metacomprehension judgments but will also be predictive of performance on tests of comprehension. Obviously, the predictive cue of disruptions in situation-level processing will not be available if individuals are not processing texts at the situation level. Therefore, the levels-of-disruption hypothesis also states that conditions that promote processing at the level of the situation model will improve the accuracy of metacomprehension judgments.<sup>2</sup>

---

<sup>1</sup>We have adopted the *levels* terminology here to connect with its use in the text-processing literature where it is typically used as a conceptual aid to distinguish between the processing of different kinds of information. When discussing various levels of representation, we do not assume spatially distinct representations, nor do we intend to imply that processing at one level excludes processing at any other. On the contrary, different kinds of information can be instantiated within a unitary representation (Kintsch, 1998), and processing while reading may occur at multiple levels. The term *levels* should also not be conflated with its use in the memory literature (e.g., in "levels of processing"), although similarities in what is meant by processing levels from the two literatures exist.

<sup>2</sup>Currently, we assume any disruptions that occur while reading have an equal chance of influencing the subsequent judgments (Rawson & Dunlosky, 2002). Other possibilities exist, however. For instance, disruptions may influence judgments only when a reader cannot resolve the processing difficulty that produced it. The level-of-disruption hypothesis concerns only one factor (i.e., disruptions) and currently does not include auxiliary assumptions about how other factors (e.g., the resolution of difficulties) moderate the effects of disruptions on metacomprehension judgments. Undoubtedly, understanding the relations among processing disruptions, online resolution of processing difficulty, and other moderating factors will be critical for a complete account of how people assess text comprehension—one that will likely be both more refined and more general than the levels-of-disruption hypothesis presented here.

Most previous investigations of metacomprehension have involved conditions that would likely not promote processing at the situation level, which may partly explain the low levels of accuracy typically reported. Some research has suggested that rereading can promote processing at the level of the situation model (Millis, Simon, & tenBroek, 1998), which suggests a straightforward prediction: Rereading texts prior to making metacomprehension judgments will improve their accuracy. Consistent with this prediction, we have demonstrated a *rereading effect* on judgment accuracy (Rawson et al., 2000): Across two experiments, the level of accuracy after rereading was approximately .60, showing a relatively substantial boost from the accuracy after a single reading. We should note that rereading does not always improve accuracy (Maki, Holder, & McGuire, 2001), which itself presents an important mystery to solve. However, in this research, our aim was to investigate mechanisms underlying the effect when it is obtained. Relevant to this endeavor, Rawson et al. (2000) empirically disconfirmed relatively uninteresting explanations for the rereading effect, such as that rereading merely increased the reliability of test scores. Nevertheless, they did not evaluate implications of the levels-of-disruption hypothesis beyond demonstrating the rereading effect itself. Accordingly, our major goal was to replicate and extend the previous work by empirically evaluating two new predictions from the levels-of-disruption hypothesis.

Concerning the first prediction, Millis et al. (1998) reported results suggesting that when two study trials are presented in relatively close succession, processing of the situation model is greater during the second reading trial than during the first. In contrast, their results also suggested that when the second reading trial is delayed by 1 week, processing during rereading is no longer primarily at the situation level but instead has shifted back to text-base processing, similar to when the texts were being read for the first time. Therefore, according to the levels-of-disruption hypothesis, when rereading is delayed for 1 week, the rereading effect on metacomprehension accuracy is not expected to occur. To evaluate this prediction, we had participants read six texts once and then either reread the texts immediately or after a 1-week delay, after which they predicted their performance on an upcoming test.

Concerning the second prediction, performance on tests that tap *comprehension* of the text rather than just memory are particularly dependent on situation-level understanding (Kintsch, 1994); hence, the rereading effect is expected to be relatively pronounced when the criterion test taps comprehension and not just memory. To evaluate this prediction, our criterion test included two kinds of questions. For each text, three questions primarily tapped memory for the text content, whereas another three questions required a deeper understanding of the content (e.g., questions requiring inferences or application of principles to new problems). The levels-of-disruption hypothesis predicts that the rereading effect will be more evident when the test consists of inference questions than when it consists of questions that focus on memory for the text-base content.

## METHOD

### Participants, Design, and Materials

Undergraduates ( $N = 113$ ) from The University of North Carolina at Greensboro participated in one of three experimental groups: single reading ( $n = 39$ ), immediate rereading ( $n = 38$ ), or delayed rereading ( $n = 36$ ). The seven texts (1 practice text and 6 critical texts) from Rawson et al. (2000, Experiment 1) were used. The texts were adapted from passages included in a Graduate Record Examination practice book. For each text, six multiple-choice questions were used: Three tapping information that was explicitly stated in the text, and three questions tapping information that could be inferred from the text. An excerpt from one of the texts (on the topic of obesity) and its corresponding questions are presented in Appendix A, and the average characteristics of all the critical texts are provided in Appendix B. Computers controlled text presentation and data collection.

### Procedure

Participants received instructions appropriate to their experimental group. All participants were instructed as follows:

The tasks you will be asked to complete today are much like those you would do when learning material for an upcoming test in one of your classes. You will be asked to read short segments of texts, judge your learning for those texts, and then later complete a test drawing on the information contained in the texts. While studying, you should do your best to understand and remember the information and ideas expressed in each text.

Therefore, their purpose for reading was to understand and remember the texts in preparation for an exam. No specific incentives were provided for task completion. After the general instructions, all participants practiced the experimental tasks with a sample text. The sample text was presented one sentence at a time (as in Maki, 1995; Maki, Jonas, & Kallod, 1994), beginning with the title. Each sentence remained on the screen until participants advanced to the next sentence with a key press. After the last sentence of each text, a performance prediction was prompted with the text title and a query: "How well do you think you will be able to answer test questions over this material in about 20 minutes? 0 (*definitely won't be able*), 20 (*20% sure I will be able*), 40 ... , 60 ... , 80 ... , 100 (*definitely will be able*)."

Participants then answered sample test questions.

For the critical trials, texts were presented in random order, one sentence at a time, as described earlier. Participants in the single reading group predicted performance for each text immediately after reading each one. After reading all texts and

making predictions, they completed the test questions and made a confidence judgment for each. Participants in the immediate rereading group first read each text once. They then reread texts in the same order, immediately predicting their performance for each after rereading. After all texts were reread and predictions were made, participants completed the test questions and made a confidence judgment for each. Participants in the delayed rereading group read each text once and were then asked to return 1 week later. On their return, participants read each text again in the same order as during the first session and immediately predicted their performance after rereading each text. After all texts were reread and predictions were made, participants completed the test questions. The time to complete the entire task was typically under 75 min, so it was unlikely that fatigue had an undue influence on participant performance.

## RESULTS

We first present the predictive accuracy of the judgments, which are critical for evaluating the two predictions from the levels-of-disruption hypothesis. We then describe an analysis of reading times (based on Millis et al., 1998), which provides evidence concerning whether rereading promoted processing at different levels of the text representation. Finally, for completeness, test performance and judgment magnitudes are presented. All differences declared as reliable have  $p < .05$ .

### Predictive Accuracy

Predictive accuracy was operationalized as the intraindividual gamma correlation between an individual's performance predictions and test performance across the six experimental texts (detailed rationale for using gamma as a measure of the relative accuracy of metacognitive judgments is provided by Nelson, 1984). Mean correlations across individual values (and corresponding medians in parentheses) were .29 (.33) for those who read once, .53 (.60) for those who reread immediately, and .22 (.33) for those who reread after a 1-week delay:  $F(2, 105) = 3.54$ ,  $MSE = .89$ ,  $p < .05$ . Participants who reread immediately were reliably more accurate than those who read once,  $t(71) = 2.01$ ; and those who reread after a 1-week delay,  $t(67) = 2.54$ . Accuracy for the latter two groups did not reliably differ,  $t(72) = .60$ . Therefore, under these conditions, a rereading effect occurred after immediate rereading (as in Rawson et al., 2000) but not after delayed rereading. The effect of immediate rereading on accuracy was quite substantial, with the accuracy of 74% of the participants who immediately reread texts being greater than the mean level of accuracy (.26) for individuals in the other two groups. Moreover, 29% of those individuals who immediately reread texts had a correlation of 1.0, whereas only 6% of those who reread after a 1-week delay achieved this level of accuracy.

For each individual, we also computed two additional gamma correlations: one between predictions and performance for the inferential questions and one between predictions and performance for the memory-based questions. Means across individual values are presented in Figure 1 along with corresponding median values (reported within each bar) and standard errors of the mean. Note that only three questions of each kind were used for each text. Because gamma correlations can be constrained when fewer questions are used (Weaver, 1990), the values based on only three questions per text (Figure 1) should not be compared to the values reported earlier, which are based on the full set of six questions per text. Although the main effect of kind of question was not reliable,  $F(1, 103) = .77$ ,  $MSE = .26$ , the main effect of group,  $F(2, 103) = 3.90$ ,  $MSE = 1.25$ , and the interaction,  $F(2, 103) = 4.04$ ,  $MSE = 1.36$ , were reliable. Follow up  $t$  tests revealed that partici-

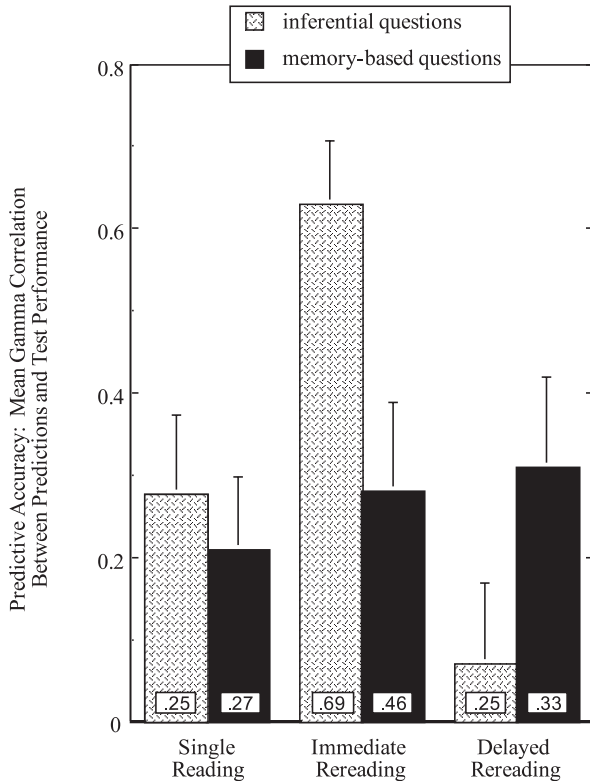


FIGURE 1 Means across individual gamma correlations between performance predictions and test performance across texts, presented as a function of group and the kind of test question. Each mean is presented with the corresponding median across individual gamma correlations (the numerical values within the graph), and error bars represent standard errors.

pants who reread immediately were more accurate on inferential questions than were those who read once,  $t(71) = 2.84$ , and those who reread after 1 week,  $t(67) = 4.55$ . Accuracy on inferential questions for the latter two groups did not differ,  $t(72) = 1.54$ . Accuracy on explicit memory-based questions did not differ between the three groups,  $t_s < .75$ , although the median values suggest that rereading also may have had a non-negligible boost on accuracy here as well. These outcomes are consistent with predictions from the levels-of-disruption hypothesis.

### Analyses of Reading Times

According to the levels-of-disruption hypothesis, rereading influences how individuals process the text, with a shift occurring across trials from more processing of the text base to more processing of the situation model. To evaluate the extent to which rereading promoted processing at different levels of the text representation, we analyzed reading times using the regression technique reported by Millis et al. (1998). The authors inferred shifts in the level of processing across reading trials by examining changes in regression coefficients when sentence reading times were regressed on several predictor variables that were theoretically related to different levels of processing. Among others, they used the following predictor variables (for justification of the use of each of these measures to indicate particular levels of processing, see Millis et al., 1998): The mean word frequency of the content words in each sentence was used as a measure of surface-level processing. The number of propositions per sentence and the number of new argument nouns (i.e., nouns introducing concepts that had not previously appeared in the text) were used as measures of text-base processing. The rated importance of each sentence to the overall meaning of the text was used as an indicator of situation-level processing, which was based on the intuitive idea that the information most central to the overall meaning of the passage also should be represented in the situation model for that text.

Among other results, Millis et al. (1998) found that indicators of text-base processing were less predictive of immediate rereading times than initial reading times. By contrast, sentence importance ratings were more predictive of rereading times than initial reading times when rereading occurred immediately, but not when rereading occurred after a 1-week delay. Based on these outcomes, Millis et al. (1998) concluded that readers shifted toward processing the situation model during immediate rereading. Although we expected the same outcome in this research, one caveat should be noted about operationalizing situation-level processing with sentence importance ratings. Namely, these ratings are unlikely pure indicators of situation-level processing and may also reflect aspects of processing at the text-base level, such as construction of the macrostructure of the text base (i.e., a representation of the main ideas in a text and their interrelations). If sentence importance ratings were unduly influenced by text-base processing, they may be less related to rereading times than to initial reading times. Therefore, this potential outcome would not necessarily indicate that rereading failed to enhance situation-level processing.

As in Millis et al. (1998), we conducted multiple regression analyses for each of our participants using the four predictor variables described earlier. We collected sentence importance ratings and computed values for predictor variables in the manner described in Millis et al. (1998). An additional 29 participants were recruited from the same population used in this experiment to collect the sentence importance ratings. The regression analysis was also conducted as in Millis et al. (1998). To trim the reading time data of any possible outliers due to aberrant responses, we first Winsorized the sentence reading time data, using a cutoff criterion of 2 standard deviations above each participant's mean (as in Millis et al., 1998). A regression analysis (across the sentences in all 6 texts) was conducted for each reading trial for each participant, and then the mean across individuals' coefficients for each predictor variable was computed for each reading trial (see Table 1). We computed both unstandardized and standardized slope coefficients; because our conclusions were identical for both analyses, we report only the former.

Two critical outcomes are evident from inspection of Table 1. First, for the immediate rereading group, the magnitude of the slope coefficients for word frequency and number of propositions decreased from the first trial to the second trial,  $t(37) > 3.10$ . These effects are specific to immediate rereading, as the comparable coefficients for the delayed rereading group changed minimally across trials,  $t(35) < 1.0$ . The decrease in slope coefficients for the new argument nouns was not statistically reliable for immediate rereaders,  $t(37) = 1.88, p = .07$ ; but was reliable for delayed rereaders,  $t(35) = 2.54$ .

Second, in contrast to expectations, the magnitude of the slope coefficients for sentence importance also decreased across trials for the immediate rereaders,  $t(37) = 2.30$ ; whereas the apparent decrease for delayed rereaders was not statistically reliable,  $t(35) = 1.64, p = .11$ . Therefore, we failed to obtain the same pattern of results as Millis et al. (1998). One possible reason for this failure can be evaluated by reanalyzing the slope coefficients separately for above average and below average performers. In particular, Millis, King, and Kim (2000) reported several conditions in which the reading level of the participants moderated processing at the situation

TABLE 1  
Predicting Sentence Reading Times: Mean Slope Coefficients for Predictor Variables Across Conditions and Reading Trials

<i>Predictor</i>	<i>Immediate Rereading</i>		<i>Delayed Rereading</i>	
	<i>First Trial</i>	<i>Second Trial</i>	<i>First Trial</i>	<i>Second Trial</i>
Word frequency	-220 (57)	50 (54)	-150 (60)	-182 (62)
Propositions	453 (30)	134 (25)	435 (22)	413 (26)
New argument nouns	117 (28)	51 (27)	129 (39)	33 (25)
Sentence importance	186 (59)	19 (51)	335 (103)	110 (82)

*Note.* Entries are means across individual unstandardized slope coefficients, with corresponding standard errors in parentheses.

level. In these cases, slope coefficients involving situation-level measures increased for above average readers and sometimes decreased for below average readers. Will the same pattern be apparent in these data? To answer this question, we used a median split based on overall test performance to distinguish readers who were below average performers from those who were above average performers. We then computed the mean coefficients from the regression analyses (described earlier) across participants in each subgroup. The corresponding outcomes are presented in Table 2. In contrast to the Ability  $\times$  Rereading interaction that might have been expected from the results of Millis et al. (2000), the pattern of outcomes was similar for above and below average performers. Most important, for immediate rereading in which we expected coefficients to increase with rereading, the coefficients involving sentence importance ratings for below and above average performers decreased almost identically across reading trials. We touch on possible explanations for these outcomes in the Discussion section.

### Test Performance and Judgment Magnitudes

Concerning test performance, the predictions from the levels-of-disruption hypothesis do not require rereading to improve test performance, especially if individuals are unable to repair the incoherence that induces a processing disruption. Readers presumably base the judgments on the amount of disruptions of processing, and these disruptions are merely more indicative of *differential* performance across texts after rereading the texts than after reading each text once. In this case, accuracy may be relatively high after rereading, even if much of the processing at this level is un-

TABLE 2  
Mean Slope Coefficients for Predictor Variables as a Function  
of Poor Versus Good Performers

Predictor	Immediate Rereading		Delayed Rereading	
	First Trial	Second Trial	First Trial	Second Trial
Below average performers <sup>a</sup>				
Word frequency	-175 (113)	69 (93)	-231 (93)	-120 (81)
Propositions	371 (50)	76 (35)	450 (30)	442 (38)
New argument nouns	119 (52)	43 (45)	120 (64)	78 (39)
Sentence importance	187 (87)	15 (86)	355 (160)	3 (100)
Above average performers <sup>b</sup>				
Word frequency	-252 (58)	35 (67)	-55 (66)	-261 (96)
Propositions	509 (33)	175 (34)	417 (33)	375 (32)
New argument nouns	115 (33)	56 (34)	139 (42)	-23 (25)
Sentence importance	186 (81)	22 (65)	312 (125)	244 (133)

*Note.* Entries are means across individual unstandardized slope coefficients, with corresponding standard errors in parentheses. Above versus below average performers were classified by a median split on test performance across all questions.

TABLE 3  
Mean Test Performance and Mean Judgment Magnitude

<i>Condition and Question Type</i>	<i>Test Performance</i>	<i>Judgment Magnitude</i>
Read once		
Overall questions	47 (2.4)	35 (2.2)
Inference questions	46 (2.7)	— <sup>a</sup>
Memory questions	48 (2.5)	—
Reread immediate		
Overall questions	53 (2.7)	42 (3.5)
Inference questions	53 (3.2)	—
Memory questions	54 (2.7)	—
Reread delay		
Overall questions	51 (2.6)	45 (3.0)
Inference questions	53 (3.1)	—
Memory questions	49 (2.8)	—

*Note.* Values are mean percentage correct test performance and mean percentage predicted test performance. Values in parentheses are standard deviations. Memory questions involved testing for recognition of explicitly stated text content.

<sup>a</sup>Predictions were not made separately for inferential questions and memory questions.

successful and, hence, does not reliably improve performance. Concerning judgment magnitudes, the between-subject design used here may reduce the likelihood of observing the effects of rereading on judgment magnitudes (cf. Carroll & Nelson, 1993). Nevertheless, although analyses of test performance and judgment magnitudes are less relevant to our present goals, we report them for consistency with previous research on text comprehension and metacomprehension. For each participant, we calculated the proportion of correct responses across the six test questions and the median performance prediction across the six texts. Means across these values are reported in Table 3.

For test performance, the rereading manipulation did not reliably influence (a) overall test performance collapsed across kind of test question; (b) test performance on inferential questions alone; or (c) test performance on explicit information questions,  $F_s < 2.0$ ,  $MSEs < 2.0$ .<sup>3</sup> The trend toward higher predictions after rereading was not reliable,  $F(2, 11) = 2.65$ ,  $MSE = 879.3$ ,  $p < .08$ .

<sup>3</sup>The absence of a performance advantage for the delayed rereading group may seem inconsistent with the spacing effects on memory for simple verbal materials that have been reported in many earlier studies. However, Schmidt and Bjork (1992) noted several studies on the spacing of practice showing that performance may be greater after massed practice than distributed practice when learning is tested immediately. Moreover, the extent to which spacing effects generalize to more complex text materials has not yet been firmly established. Although some research has demonstrated greater memory for text content after distributed versus massed rereading (Glover & Corkill, 1987; Krug, Davis, & Glover, 1990), more recent research has shown the opposite pattern (Rawson & Kintsch, 2004). This research provides further evidence that spacing effects with text material may not be as robust as previously assumed.

## DISCUSSION

A historical review of metacomprehension research reveals that effects on predictive accuracy are not common in this literature (for a detailed review, see Weaver et al., 1995). Research conducted in the 1980s led to the conclusion that people's metacomprehension accuracy was low to nonexistent. Weaver (1990) demonstrated that unreliable measurement was partly to blame by showing that the format of criterion tests artifactually limited estimates of predictive accuracy. Even when effects were evident, however, predictive accuracy was still relatively low. For instance, Maki, Foley, Kajer, Thompson, and Willert (1990) reported that deleting letters from words in a text improved predictive accuracy, but mean accuracy for the deleted-letters text was lower than .40.

In contrast to the pessimism warranted by earlier research, Weaver and Bryant (1995), Thiede et al. (2003), and Rawson et al. (2000) recently discovered variables that can have relatively substantial effects on accuracy. First, Weaver and Bryant (Experiment 2) reported what we refer to as the *text-difficulty* effect: Mean predictive accuracy was greater for texts of moderate difficulty ( $M$  gammas  $\approx .75$ ) than for either easier ( $M \approx .20$ ) texts or more difficult ( $M \approx .33$ ) texts. Second, Thiede et al. had college students read texts on various topics. Some time after reading each one, a participant was presented with the title of the text and asked to write five keywords that captured the essence of its content, and then asked to make a metacomprehension judgment. Generating keywords substantially boosted accuracy when keywords were generated at a delay after all texts had been read versus when they had been generated immediately after reading each text. Finally, Rawson et al. observed the rereading effect in two experiments independently conducted at different universities. The methods employed at the two sites also differed on numerous dimensions, including the difficulty of the texts, the number of texts, the prompt for metacomprehension judgments, and the method of presenting texts for reading. Despite these differences, both methods yielded a rereading effect (however, see Maki et al., 2001, for a failure to replicate).

Two new outcomes from this research constrain hypotheses of the rereading effect as well as general theories of metacomprehension accuracy: The rereading effect was most prominent when learning was measured by questions requiring comprehension of the text content, and the effect was absent when rereading occurred after a 1-week delay. Although these outcomes were consistent with predictions derived from the levels-of-disruption hypothesis, two issues concerning this hypothesis should be considered.

First, one may argue that the levels-of-disruption hypothesis should predict a *decline* in accuracy after rereading. More specifically, if an individual monitors disruptions of text processing, he or she may devote more effort to comprehending the incoherent portions of the text, which may yield lower judgments (due to

disruptions) but better comprehension (due to increased effort). In this case, basing judgments on disruptions would lower predictive accuracy. Although plausible, the effects of increased effort on comprehension may be rather limited. Namely, effortful processing may improve comprehension of incoherent text, but the benefits of extra effort appear limited to readers who have the appropriate background knowledge to repair the incoherence (McNamara, Kintsch, Songer, & Kintsch, 1996). For less knowledgeable students, extra effort presumably would not improve performance on a comprehension test. An intriguing implication is that the knowledge level of readers will moderate the rereading effect, with a counterintuitive prediction being that the effect will be smaller for high-knowledge readers than for low-knowledge readers.

The second issue concerns a critical assumption of the levels-of-disruption hypothesis, namely, that the reason rereading improves accuracy is because it increases the likelihood readers will process at the situation level for the texts. However, did rereading increase the likelihood of such processing? Evidence from the regression analysis at first suggests that rereading did not have the expected effect. In particular, with immediate rereading, reading times were expected to be less highly related to measures of text characteristics indicative of the text base and more highly related to ratings of sentence importance (the predictor of situation-level processing). Although the former outcome was evident, the latter was not. Although the latter outcome does pose problems for the levels-of-disruption hypothesis, alternative interpretations of the reading time data need to be considered. We offer two alternatives here. First, subjective ratings of sentence importance may be more indicative of processing the macrostructure of the text-base representation than the situation model. The macrostructure involves the representation of the topical structure of a text (Kintsch, 1998) but does not necessarily include the integration of text content with prior knowledge that presumably underlies the development of a situation model. If so, the importance ratings may provide another measure of processing at the level of the text base rather than of the situation model.

The second alternative is based on one speculation offered by Millis et al. (2000) about why below average readers sometimes do not show expected increases in the coefficients for sentence importance ratings; a speculation that may hold for both ability groups in this context: "One explanation for this finding is that the below average readers were overconfident of the knowledge contained in the text . . . . As a consequence, they did not bother constructing the situation model" (p. 231). In this context, perhaps participants were in general overconfident in their understanding of the text after an initial reading, which in turn reduced their efforts while rereading. Outcomes from Table 3 are relevant to evaluating this possibility. On average, participants judged that they would correctly answer only about 35% of the questions, whereas their overall test performance was 47%, which demonstrates 12% underconfidence after the initial reading trial. Therefore, although overconfidence may have reduced situation-level

processing in the experiments described by Millis et al. (2000), it apparently was not influential in this experiment.

Although the levels-of-disruption hypothesis did receive empirical support in this research, other hypotheses provide competing explanations that should be considered. We briefly discuss two hypotheses that hold promise for understanding the rereading effects as well as other effects that have been reported in the literature.

The first is the *optimum-effort* hypothesis, which was proposed by Weaver and Bryant (1995) to account for the effect of text difficulty on accuracy but is also germane to rereading effects. According to this hypothesis, when text processing is too effortful, such as with texts that exceed participants' reading level, the effort required to process the text may compromise their monitoring of comprehension; this, in turn, will constrain accuracy. When text processing is too easy, readers may not be sufficiently engaged by the text, which "drives the reader into an 'automatic' reading mode" (Weaver & Bryant, 1995, p. 19) that may not trigger online monitoring. Finally, when text processing is optimally effortful, such as with texts of moderate difficulty that match the participants' reading level, participants may be more engaged by the text and will expend more effort in monitoring comprehension; this, in turn, leads to higher levels of accuracy. Presumably, the optimum level of effort for processing and monitoring will most likely be achieved when reader ability and text difficulty match. The optimum-effort hypothesis can also account for the rereading effect, assuming that rereading increases the match between reader ability and the demands of text processing in a way that supports better monitoring during rereading. Outcomes from Lin, Zabrocky, and Moore (1998), however, are inconsistent with one prediction from this optimum-effort hypothesis. In particular, they found that accuracy was higher for moderately difficult texts (as compared to the more difficult texts) even for individuals with reading levels that best matched the more difficult texts. Therefore, the match between reading level and text difficulty level *per se* may be less important to accuracy than is suggested by the optimum-effort hypothesis.

A more general *resource-availability* hypothesis can account for the extant data. According to this hypothesis, accuracy will be higher in conditions in which individuals have more (vs. less) available resources to allocate to monitoring online comprehension. To explain the text difficulty effects mentioned earlier, perhaps the most difficult texts are so difficult that even individuals with the highest reading levels would have little resources available to monitor comprehension during an initial reading. In addition, the finding that accuracy is also low for very easy texts (Weaver & Bryant, 1995) is not necessarily inconsistent with the hypothesis. With easy texts, accuracy could be constrained because variability in difficulty across a group of easy texts may be attenuated—that is, resources for monitoring may be available, yet the homogeneity of the easy texts may undermine readers'

ability to discriminate between them. Finally, with respect to how the resource-availability hypothesis might account for the rereading effect, previous studies have demonstrated that more cognitive resources are available during rereading than during an initial reading (e.g., Inhoff & Fleming, 1989; Levy, Di Persio, & Hollingshead, 1992; but, see Raney, 1993), which suggests that rereading may afford more resources for comprehension monitoring.

This resource-availability hypothesis differs qualitatively from the levels-of-disruption hypothesis. With regard to the rereading effect, the latter hypothesis implies that the extent to which readers are monitoring during reading does not differ during the first versus second reading trial, but instead that the cues available as a basis for monitoring are merely more predictive of test performance after rereading than after an initial reading. By contrast, the resource-availability hypothesis implies that people are better able to monitor comprehension while rereading because more resources are available for monitoring. These hypotheses could be competitively evaluated with respect to explaining metacomprehension accuracy (i.e., some manipulations presumably would influence the cognitive resources demanded by text processing without influencing processing at the situation level, or, more generally, without influencing the level at which texts are processed). In these cases, the resource-availability hypothesis predicts that manipulation will influence accuracy, whereas the levels-of-disruption hypothesis does not.

A final point concerns the practical implications of the rereading effect we have demonstrated here and in earlier research. We are wary of prescribing rereading as a *panacea* for improving students' metacomprehension accuracy for several reasons: (a) As demonstrated in this research, rereading after a moderate delay does not improve predictive accuracy; (b) assuming that availability of cognitive resources are vital for improving accuracy, even immediate rereading may not consistently improve predictive accuracy; and perhaps most important, (c) even when rereading does improve accuracy, it is still substantially below perfect (i.e., accuracy was just above +.50 in this research). Many theoretical challenges remain before researchers will be able to prescribe how students can consistently obtain high levels of predictive accuracy for text material. Toward meeting these challenges, however, the discovery of manipulations that moderate metacomprehension accuracy, such as those described earlier, bodes well for the future success of understanding the processes underlying metacomprehension accuracy.

#### ACKNOWLEDGMENTS

John Dunlosky and Katherine Rawson are now at Kent State University, Department of Psychology.

## REFERENCES

- Carroll, M., & Nelson, T. O. (1993). Effects of overlearning on the feeling of knowing are more detectable in within-subject than in between-subject designs. *American Journal of Psychology*, *106*, 227–235.
- Glover, J. A., & Corkill, A. J. (1987). Influence of paraphrased repetitions on the spacing effect. *Journal of Educational Psychology*, *79*, 198–199.
- Hacker, D. J. (1998). Self-regulated comprehension during normal reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 165–191). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Inhoff, A. W., & Fleming, K. (1989). Probe-detection times during the reading of easy and difficult texts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 339–351.
- Kintsch, W. (1994). Learning from text. *American Psychologist*, *49*, 294–303.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.
- 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.
- Krug, D., Davis, B., & Glover, J. A. (1990). Massed versus distributed repeated reading: A case of forgetting helping recall? *Journal of Educational Psychology*, *82*, 366–371.
- Levy, B. A., Di Persio, R. D., & Hollingshead, A. (1992). Fluent rereading: Repetition, automaticity, and discrepancy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 957–971.
- Lin, L. M., Zabrocky, K., & Moore, D. (1998). Effects of text difficulty and adults' age on relative calibration of comprehension. *American Journal of Psychology*, *115*, 187–198.
- Maki, R. H. (1995). Accuracy of metacomprehension judgments for questions of varying importance levels. *American Journal of Psychology*, *108*, 327–344.
- Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–144). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 609–616.
- Maki, R. H., Holder, E. W., & McGuire, M. J. (2001, November). *Metacomprehension of text: A test of the optimum effort hypothesis*. Paper presented at the 42nd Annual Meeting of the Psychonomics Society, Orlando, FL.
- Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review*, *1*, 126–129.
- 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.
- Millis, K. K., King, A., & Kim, H. J. (2000). Updating situation models from descriptive texts: A test of the situational operator model. *Discourse Processes*, *30*, 201–236.
- Millis, K. K., Simon, S., & tenBroek, N. S. (1998). Resource allocation during the rereading of scientific texts. *Memory & Cognition*, *26*, 232–246.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–133.
- Raney, G. E. (1993). Monitoring changes in cognitive load during reading: An event-related brain potential and reaction time analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 51–69.

- Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *XX*, xxx–xxx.
- Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, *28*, 1004–1010.
- Rawson, K. A., & Kintsch, W. (2004). *Rereading effects depend upon time of test*. Manuscript submitted for publication.
- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.
- Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204–221.
- Thiede, K. W., Anderson, M. C. M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of text. *Journal of Educational Psychology*, *95*, 66–73.
- Weaver, C. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 214–222.
- Weaver, C. A., III, & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition*, *23*, 12–22.
- Weaver, C. A., III, Bryant, D. S., & Burns, K. D. (1995). Comprehension monitoring: Extensions of the Kintsch and van Dijk model. In C. A. Weaver, III, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 177–193). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

## APPENDIX A

The following is an excerpt from a critical text on the topic of obesity along with the corresponding inference-based and memory-based test questions (asterisks appear beside correct answers but did not appear on the actual test, and the order of alternatives was randomized anew for each participant). Characteristics of each of the six critical texts are presented in Appendix B. The complete set of texts and corresponding questions are available from John Dunlosky.

### Obesity

In terms of its prevalence, obesity is the leading disease in the United States. Obesity may be defined as a condition of excess adipose tissue, as fatness beyond cultural esthetic norms, or as adipose tissue tending to disrupt good health of mind and body. A common rule of thumb is that people more than 20 pounds above their desirable weight are obese; by this measure, 30% of men and 40% of women in America are obese. Despite the prevalence of the disease, curative measures are almost impossible for those currently obese; future generations may be spared. Adipose tissue is a triumph of evolution. Fat yields 9.0 calories per gram, while carbohydrates and protein each yield 4.0 calories per gram, and fat contains much less water than does protein. It is, therefore, much more efficient to store excess energy as fat than as protein ... .

### Inference-Based Questions

It can be inferred from the passage that

- the roots of obesity are to be found in the feeding and eating problems of infancy and childhood (\*)
- following a careful weight-loss diet is the only effective cure for obesity
- bringing the body into a condition of negative nitrogen balance will assist the dieter in achieving weight loss
- atherosclerotic people also suffer from obesity
- psychiatric treatment can uncover the underlying causes of obesity

For which of the following questions does the passage suggest an answer through the information it offers?

- Why do people very often fail to lose weight even though they are cutting down on caloric intake? (\*)
- How do hypertension and atherosclerosis contribute to obesity in modern man?
- What is the effect of insulin on metabolism and weight loss?
- How can obese parents remedy weight control problems in their children?
- What part does metabolic rate play in the utilization of carbohydrates to accelerate weight loss?

According to the passage, the role that evolution plays in relation to obesity is that

- modern man's body cannot deal with an evolutionary vestige of what was needed by primitive man, the development of adipose tissue for storing energy in the face of uncertain food supply (\*)
- adipose tissue is a convenient form of body structure in which to store excess protein
- modern man uses large amounts of energy, mostly in the form of protein and carbohydrates
- the development of a sedentary lifestyle encouraged the ingestion of reduced calories
- primitive man's need for insulation from the cold led to modern man's need for a diet strong in serum lipids

### Memory-Based Questions

According to the statistics presented in the passage,

- 30% of men and 40% of women in America are obese (\*)
- 40% of men and 30% of women in America are obese

- c. 20% of men and 40% of women in America are obese
- d. 20% of men and 30% of women in America are obese
- e. 30% of men and 30% of women in America are obese

The leading cause of death in the United States is

- a. atherosclerotic heart disease (\*)
- b. hyperinsulinemia
- c. hypertension
- d. diabetes
- e. obesity

Which of the following statements is NOT true?

- a. carbohydrates yield more calories per gram than protein (\*)
- b. protein yields four calories per gram
- c. fat contains less water than protein
- d. fat yields nine calories per gram
- e. carbohydrates yield fewer calories per gram than fat

## APPENDIX B

TABLE B1  
Text Characteristics

<i>Text Title</i>	<i>F Reading Ease</i>	<i>F-K Level</i>	<i>Number Sent</i>	<i>Words/Sent</i>	<i>Test Performance</i>			
					<i>Overall</i>	<i>Inference</i>	<i>Memory</i>	<i>Judgment</i>
Guilt	46.4	11.2	17	21.1	52 (2.4)	58 (3.2)	45 (2.6)	47 (2.0)
Intelligence	27.7	14.7	21	24.8	38 (2.1)	40 (2.7)	35 (2.7)	29 (1.7)
Inventions	21.1	17.4	16	28.4	45 (2.0)	36 (2.6)	53 (2.8)	32 (2.0)
Literature	47.6	11.4	25	22.4	55 (2.3)	52 (3.2)	58 (2.6)	45 (2.2)
Majority	41.9	12.6	16	23.7	55 (2.2)	58 (2.8)	51 (2.8)	36 (2.1)
Obesity	44.3	12.6	23	26.1	61 (2.3)	63 (3.0)	59 (2.9)	56 (2.0)

*Note.* F = Flesch; F-K = Flesch-Kincaid; Number Sent = number of sentences per text. Words/Sent = mean number of words per sent. Memory questions involved testing for recognition of explicitly stated text content. Judgment is the mean judgment magnitude across participants for a particular text. Entries in parentheses are standard errors of the corresponding values across participants.