Memory & Cognition
2001, 29 (1), 62-67

The postdiction superiority effect in metacomprehension of text

BENTON H. PIERCE and STEVEN M. SMITH
Texas A&M University, College Station, Texas

Metacomprehension accuracy for texts was greater after, rather than before, answering test questions about the texts—a postdiction superiority effect. Although postdiction superiority was found across successive sets of test questions and across successive texts, there was no improvement in metacomprehension accuracy after participants had taken more tests. Neither prediction nor postdiction gamma correlations with test performance improved with successive tests. Although the results are consistent with retrieval hypotheses, they contradict predictions made by test knowledge hypotheses, which state that increasing knowledge of the nature of the tests should increase metacomprehension accuracy.

Glenberg and Epstein (1985) observed that although predictions of performance on tests of text comprehension are often poor, retrospective assessments of test performance, also known as *postdictions*, are more accurate. The correlation between confidence in the correctness of answers and actual test performance has been called *calibration of performance* (see, e.g., Glenberg & Epstein, 1987; Glenberg, Sanocki, Epstein, & Morris, 1987); better calibration refers to more accurate estimates of test performance. We will refer to the accuracy of postdiction relative to prediction judgments as the *postdiction superiority effect.*

Maki and Serra (1992) provided a clear example of a postdiction superiority effect in metacomprehension of text. Participants made predictions and postdiction judgments about their performance on tests of reading comprehension. Prereading judgments were given before reading texts, based only on topic familiarity of the texts. Postreading judgments were made after reading the texts, but before answering questions about the text. Posttest judgments were made after answering four test questions. Thus, in Maki and Serra's study, both prereading and postreading judgments were predictions of later performance, whereas posttest judgments were postdictions. Correlations between judgments and performance were higher for the posttest phase than for either the prereading or the postreading phases, a postdiction superiority effect.

In the present study, we explore two classes of hypotheses to account for the reported postdiction superiority effects: *retrieval* hypotheses and *test knowledge* hypotheses (see Table 1). Retrieval hypotheses state that postdiction involves retrieving what happened when specific test questions were asked. For example, one version of the class of retrieval hypotheses is derived from the Glenberg et al. (1987) suggestion that participants can use self-generated feedback from answering questions to judge their confidence that answers are correct. A participant, for instance, who remembers having given what seemed like satisfactory answers to three of the four test questions would postdict that three questions were answered correctly. Another example of a retrieval hypothesis states that postdictions may be based on how plausible or difficult the participant remembers the distractors on those items to have been. In either case, the participant's assessment of prior performance is based on retrieval of events related to already experienced test items. According to this class of hypotheses, postdiction judgments should be consistently more accurate than prediction judgments because memories of answering the questions are unavailable before one takes the test.

Alternatively, Maki (1998b) suggested that increased exposure to test questions may cause increases in metacomprehension accuracy as participants gain additional information about the nature of the tests. Test knowledge hypotheses are concerned with the participant's assessment of the general difficulty of the criterion test itself. One test knowledge hypothesis, for example, proposes that one's assessment of test difficulty is based on learning whether the questions test verbatim memory or one's ability to draw inferences from the test. Another example of a test knowledge hypothesis states that participants draw on experience from having taken prior tests in predicting how difficult or picky the test questions are likely to be. In either case, test knowledge is acquired from having taken prior tests. According to these hypotheses, accuracy should become progressively better as more tests are taken, because participants accumulate progressively greater amounts of test knowledge. Thus, test knowledge hypotheses predict that until the participant is extremely

We thank Deborah Bryant and Charles Weaver for kindly providing us the texts and for their helpful comments concerning this research. We also thank Jenny Acklam and Jason Warrington for helping score data. Address all correspondence to B. H. Pierce, Department of Psychology, Texas A&M University, College Station, TX 77843-4235 (e-mail: bhp6960@acs.tamu.edu).

Table 1
Examples of Retrieval and Test Knowledge Hypotheses

Retrieval Hypotheses: Remembering what happened when questions were answered.
Participant *remembers*:
1. How many answers were given
2. How picky the questions were
3. How similar the distractors were to each other
4. How confident he/she felt when answering questions
5. How fluent he/she was in processing the alternatives
Test Knowledge Hypotheses: Inferring the nature of tests.
Participant *infers*:
1. How difficult questions are likely to be
2. How specific or general questions are likely to be
3. Whether questions test facts, names, inferences, or gist

familiar with the types of test questions to be asked, metacomprehension accuracy will continue to improve.

Although the two classes of hypotheses may share some underlying mechanisms, it is nonetheless possible to tease apart different predictions of the two. Specifically, the two types of hypotheses project different outcomes in relation to postdiction superiority as the participant continues to encounter and learn more about the test questions. Retrieval hypotheses predict that postdictions will be consistently more accurate than predictions, because no matter how many tests participants have taken in the past, they cannot generate implicit feedback for a specific set of questions until after they have answered the questions. Test knowledge hypotheses, on the other hand, state that postdictions will be more accurate than predictions only until one has finished learning about the general nature and difficulty of the questions. Once participants gain enough test knowledge, predictions should be as accurate as postdictions. Therefore, as participants encounter more test questions and gain test knowledge, the postdiction superiority effect should be attenuated.

We tested these two classes of hypotheses in two experiments in which participants read multiple texts, with each text followed by 16 test questions over the text material. The 16 questions for each text were divided into four sets, with 4 questions in each set. Before each 4-question set, participants were asked to predict their performance on the upcoming set, and after each set they were asked to assess how well they thought they had done on the previous 4 questions. Both the test knowledge and retrieval hypotheses predicted a postdiction superiority effect; that is, both predicted that the gamma correlation between metacomprehension judgments and test performance would be greater for postdictions than for predictions. What distinguishes the two classes of hypotheses are their predictions concerning the interaction of the effect with earlier versus later texts and question sets.

Test knowledge hypotheses predicted that the postdiction superiority effect would diminish from the first question set through the fourth set for each text, and that the effect should also decrease from earlier texts to later ones. As participants progress from one question set to the next, and from one text to the next, they should acquire test knowledge, which should attenuate the postdiction superiority effect. If adequate test knowledge is acquired before the last text or the last question set for a text, and if test knowledge causes postdiction superiority, then the effect should not be seen for later texts and question sets.

Retrieval hypotheses predicted that postdiction superiority should be the same for all texts and question sets. These hypotheses state that information upon which to base an accurate postdiction, regardless of the state of one's test knowledge, can occur only after one has seen the questions. Therefore, postdiction superiority should occur for all question sets of all texts. If both test knowledge and information retrieved from having answered the questions contribute to postdiction superiority effects, the observed effect should diminish across texts and question sets, but it should not disappear.

It is important to note that this study cannot differentiate among the different hypotheses within a class. For example, one type of retrieval hypothesis states that postdictions are based on remembering self-generated feedback from having given what appears to be the correct answer. Another proposes that postdictions are based on remembering the relative difficulty of the distractors. In the present study, we employ no systematic method that would allow us to distinguish between these two specific hypotheses. Our more modest goal is to distinguish between two types of mechanisms that may underlie the postdiction superiority effect observed in prior studies, one based on retrieval of events that occurred when one answered the test questions, and the other based on participants' perceptions of test difficulty.

The methods and the materials used in the present study resemble those of previous metacomprehension studies, such as Maki and Serra (1992) and Weaver and Bryant (1995). The texts and text questions were drawn from Weaver and Bryant's study, although those researchers did not examine postdiction effects. The present procedure differed from Maki and Serra's study in the numbers of questions asked per text. Whereas Maki and Serra asked only 4 questions per text, we asked 16 questions per text, or four sets of 4 questions. Thus, the postdiction superiority effect observed by Maki and Serra (comparing postreading accuracy with posttest accuracy) was equivalent to the effect that we observed within the first of the four question sets.

EXPERIMENT 1

In Experiment 1 participants read and answered questions about four narrative texts taken from the materials of Weaver and Bryant (1995). The texts were simple ones, and the test questions were fairly easy. Sixteen questions, arranged in four sets of 4 questions apiece, were given after the four texts. Before and after each question set, participants were asked to judge how many of the four they would answer (or had answered) correctly. It was predicted that a postdiction superiority effect would be observed—that is, that gamma correlations relating post-

dictions with accuracy measures would be greater than correlations of predictions with accuracy. The class of test knowledge hypotheses further predicted that the postdiction superiority effect would decrease from the first to the fourth text, and from the first to the fourth question set for each text.

Method

Participants. A total of 44 participants took part in Experiment 1. These participants were Texas A&M introductory psychology students who participated in exchange for partial fulfillment of the course requirement.

Materials. The texts and questions were taken from the Weaver and Bryant (1995) study. These texts were of sufficient length so that multiple sets of questions could be derived from each text. The four narrative texts were brief fairy tales: "Old Sultan," "The Wolf and the Seven Young Kids," "The Queen Bee," and "The Owl." There were 16 questions per text, with the questions arranged in four sets of 4 questions each. Each text was divided into four sections of approximately equal length, with each set of questions referring to a specific section of the text. Altogether, each participant answered a total of 64 questions. The text passages and questions were in the form of written handouts.

Design. Texts and question sets were counterbalanced with a Graeco-Latin Square technique that allowed presentation of four unique text and question set orders. Text position (i.e., whether a text was read first, second, third, or fourth), question set (i.e., first through fourth), and type of judgment (prediction vs. postdiction) were within-subjects factors. Text and question set orders were counterbalanced between subjects.

Procedure. The same procedure was used in Experiments 1 and 2 and is shown in Table 2. Participants read all four texts in a predetermined order. Following completion of the fourth text, participants were asked to make prediction judgments regarding the criterion test for the first text. The instructions stated, "You will now be asked 16 questions about [the name of Text 1]. The first set of questions will be asked on the next page. How many of these four questions do you think you will answer correctly?" Participants then answered the four questions. They were then asked to make a postdiction judgment of how many questions they thought they had answered correctly. Then a prediction for the next set of questions was elicited, participants answered the questions, and made a postdiction for that set. Participants repeated this procedure until all four sets of questions had been answered. This procedure was repeated for all four criterion tests.

To summarize, participants read all four texts, and then made predictions and answered questions presented in blocked form. Both predictions and criterion tests, therefore, were delayed (see Maki, 1998a).

Results

Metamemory accuracy.[1] Gamma correlations relating metacomprehension judgments (which ranged from zero to four) to accuracy (maximum score was four correct) were calculated for the 16 prediction/criterion performance pairs and the 16 postdiction/criterion performance pairs. The mean gamma correlation for prediction accuracy was $r = .049$, whereas the gamma correlation for postdiction accuracy was $r = .466$. A one-way analysis of variance (ANOVA) was computed, with type of judgment (prediction vs. postdiction) as the repeated measure. The analysis showed that the postdiction superiority effect was significant [$F(1,38) = 28.18$, $MS_e = .120$, $p < .001$]; postdiction accuracy, as measured by gamma, was greater than prediction accuracy.

Gamma correlations were calculated for each participant by question set, collapsed across all four texts. These gamma correlations are summarized in Table 3. Single sample *t* tests revealed that all four of the postdiction gamma correlations reliably differed from zero, whereas none of the prediction gamma correlations differed from zero.

A one-way ANOVA was computed to compare gamma correlations across the four question set positions for prediction judgments.[2] The effect of question set position was not significant [$F(3,39) = .256$, $MS_e = .482$, $p = .856$]. Another one-way ANOVA was computed to compare gamma correlations across the four question set positions for postdiction judgments. No effect was found on postdiction gamma correlations [$F(3,75) = .688$, $MS_e = .422$, $p = .562$]. Neither prediction nor postdiction metamemory accuracy changed as participants progressed through the successive question sets.

We also examined whether metamemory accuracy, as measured by gamma correlations, changed as participants progressed through the four successive text positions. Mean prediction and postdiction gamma correlations for each of the four text positions are shown in Table 4.

Single sample *t* tests revealed that the postdiction gamma correlations for each text position reliably differed from zero, whereas the correlation involving predictions of performance on Text 4 was the only prediction gamma that was different from zero.

A one-way ANOVA was computed to compare gamma correlations across the four text positions for prediction judgments. The effect of text position was not significant [$F(3,18) = 1.84$, $MS_e = .787$, $p = .18$]. Another one-way ANOVA was computed to compare gamma correlations across the four text positions for postdiction judgments. Likewise, no effect was found on postdiction gamma correlations [$F(3,66) = .865$, $MS_e = .454$, $p = .464$].

Recognition memory performance. The mean proportion correct on the recognition test was .722.[3] Table 5 shows recognition performance for each text position and text selection.

A repeated measures ANOVA revealed no main effect of texts [$F(3,264) = 2.04$, $p > .10$]. Recognition performance on the text selections was analyzed in another repeated measures ANOVA. As expected, a main effect was found for the text selections [$F(3,264) = 21.2$, $p < .001$]. These differences in criterion performance were expected, because the texts have been found to vary in terms of readability (Weaver & Bryant, 1995).

Table 2
General Method for Experiments 1 and 2

1. Participants read all texts.
2. Participants predict performance on the first set of four questions for text 1 (i.e., they predict the number they will get correct out of four).
3. Participants answer the first question set and judge how many they answered correctly.
4. Repeat Steps 1–3 for Question Sets 2, 3, and 4.
5. Repeat Steps 1–4 for tests covering remaining texts.

Table 3
Mean Gamma Correlations for Prediction and Postdiction Judgments as a Function of Question Set Position in Experiments 1 and 2

	Question Set Position							
	Set 1		Set 2		Set 3		Set 4	
Judgment	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
				Experiment 1				
Predictions	.115	.151	.174	.159	.190	.159	.214	.141
Postdictions	.488*	.115	.656*	.089	.466*	.119	. 353*	.126
				Experiment 2				
Predictions	.355*	.151	.161	.159	.438*	.159	.306*	.141
Postdictions	. 373*	.115	.643*	.089	.577*	.119	.657*	.126

*Correlations significantly different from zero.

Discussion

A robust postdiction superiority effect was found in Experiment 1. This result is consistent with that of Maki and Serra (1992). Because Maki and Serra asked only four questions per text, their postdiction superiority effect is analogous to that found for Question Set 1 in Experiment 1.

Participants in Experiment 1 did not improve in either prediction or postdiction accuracy as they progressed from one question set to the next, or from one text to the next. Participants apparently did not use (or were unable to use) increased exposure to test questions and the information gained from answering the questions to make more accurate performance judgments on succeeding questions. These findings are consistent with retrieval hypotheses, but not test knowledge hypotheses. If participants' knowledge about the test questions increased across question sets or texts, there was no evidence of that increase in the patterns of prediction or postdiction gamma correlations. This finding does not indicate that improved test knowledge never causes postdiction superiority; it shows that the robust effect found in Experiment 1 cannot be accounted for by test knowledge hypotheses. The results do not contradict retrieval hypotheses, however. Participants appear to have been able to use retrieved information from the text pertaining to the specific items to make accurate postdiction judgments.

EXPERIMENT 2

Recognition performance in Experiment 1 was relatively high ($M = .72$), which may have reduced variability in both judgments and criterion responses. This lack of variability, in turn, could have affected the level and pattern of the gamma correlations. In addition, only narrative texts were used in Experiment 1. Weaver and Bryant (1995) found that type of text (i.e., narrative vs. expository) and associated criterion performance can affect gamma correlations. In Experiment 2, for generality, we used expository texts to test the test knowledge and retrieval classes of hypotheses. The three expository texts used in Experiment 2 were expected to produce lower criterion performance and, hence, greater variability in judgments and criterion responses.

Method

Participants. Forty-nine participants took part in Experiment 2. All were Texas A&M introductory psychology students who participated in exchange for partial fulfillment of the course requirement. Data were discarded from 2 participants who failed to make all of the metacognitive judgments. Data from the remaining 47 participants were used in Experiment 2.

Design. The design of Experiment 2 was the same as that for Experiment 1, except that only three texts were used in Experiment 2, whereas four had been used in Experiment 1. Thus, there were four question sets for each of three text positions, with both variables manipulated within subjects.

Materials. As in Experiment 1, the texts and questions were taken from Weaver and Bryant (1995). The three expository texts used were entitled "The Martian Atmosphere," "Symbiosis," and "Euripides." There were four sets of 4 questions for each text, a total of 16 questions per text. Like the narrative texts used in Experiment 1, the expository texts were divided into four sections of approximately equal length, with each set of questions referring to a specific section of the text. Participants answered 48 questions, with the texts and questions presented in written handouts.

Procedure. The procedure for Experiment 2 was the same as that of Experiment 1. Text and question set orders were counterbalanced between subjects.

Results

Metamemory accuracy. Gamma correlations were calculated for the 12 prediction/performance pairs and the 12 postdiction/performance pairs. The mean gamma correlation for prediction accuracy was $r = .282$, and that for postdiction accuracy, $r = .526$. A one-way ANOVA to compare prediction versus postdiction accuracy revealed a significant postdiction superiority effect [$F(1,42) = 9.70, MS_e = .132, p < .01$]; postdiction accuracy, as measured by gamma, was greater than prediction accuracy.

Gamma correlations were calculated for question set positions, collapsed across all three texts. These correlations are shown in Table 3. Single sample *t* tests indicated that all of the gamma correlations differed significantly from zero, with the exception of the prediction gamma for Question Set 2. Repeated measures ANOVAs with question sets as the within-subjects factor revealed no main effect for the prediction gamma correlations [$F(3,57) = .643, MS_e = .764, p = .591$], nor was there a main effect for the postdiction gamma correlations [$F(3,75) = 1.27, MS_e = .506, p = .29$], indicating no increase in prediction

Table 4
Mean Gamma Correlations for Prediction and Postdiction Judgments as a Function of Text Position in Experiments 1 and 2

	Text Position							
	Text 1		Text 2		Text 3		Text 4	
Judgment	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
				Experiment 1				
Predictions	.136	.163	−.189	.192	−.152	.172	−.435*	.177
Postdictions	.337*	.121	.363*	.089	.405*	.117	.578*	.108
				Experiment 2				
Predictions	.010	.147	−.200	.209	.229	.167		
Postdictions	.411*	.112	.313*	.115	.345*	.128		

*Correlations significantly different from zero.

Table 5
Mean Recognition Performance as a Function of Text Position and Text Selection in Experiment 1

Performance									
Text 1		Text 2		Text 3		Text 4		All Texts	
M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Text Position									
.736	.022	.703	.024	.722	.026	.726	.024	.722	.017
Text Selection									
.798	.018	.747	.020	.699	.025	.642	.025	.722	.017

Note—Text 1, "Old Sultan"; Text 2, "The Wolf and the Seven Young Kids"; Text 3, "The Owl"; Text 4, "The Queen Bee."

or postdiction accuracy as participants progressed through the question sets.

We also examined whether metamemory accuracy, as measured by gamma correlations, changed as participants progressed through the three text positions. Mean prediction and postdiction gamma correlations for the three text positions are shown in Table 4. Single-sample *t* tests showed that all three of the postdiction gamma correlations differed from zero, whereas none of the prediction gamma correlations did. One-way ANOVAs revealed no effect of text position on prediction judgments [$F(2,12) = .230$, $MS_e = 1.03$, $p = .798$] or postdiction judgments [$F(2,60) = .252$, $MS_e = .581$, $p = .778$], indicating that metamemory accuracy did not change from one text to the next.

Recognition memory performance. The mean proportion correct on the recognition test was .50; it was .72 in Experiment 1. Thus, the use of expository texts in Experiment 2 rather than the narrative texts used in Experiment 1 had the desired effect of lowering criterion performance.

These mean proportion correct scores were analyzed in a single-factor repeated measures ANOVA with the three text positions as the within-subjects factor. No effect of text position was found [$F(2,92) = 2.29, p > .10$].

A similar repeated measures ANOVA of recognition performance on the text selections revealed a main effect of text selections [$F(2,92) = 42.76$, $MS_e = .02$, $p < .001$]. Table 6 shows recognition performance for each text position and text selection.

Discussion

The results from Experiment 2, in which expository texts rather than narrative texts were used, show similar patterns to those found in Experiment 1. No increases in metamemory accuracy were detected as participants moved through the criterion tests (i.e., through successive question sets), nor did their accuracy improve as they moved through successive texts. This pattern of results clearly indicates that participants did not use increasing amounts of test knowledge to improve their subsequent metamemory accuracy.

A robust postdiction superiority effect was found. As measured by gamma, therefore, participants were clearly able to accurately assess their prior performance on the questions, owing likely to information that was retrieved surrounding the events when participants answered the questions.

The results of Experiment 2 support the class of retrieval hypotheses, but provide no support for test knowledge hypotheses. Participants were able to make more accurate retrospective or postdiction judgments than prediction judgments but apparently did not use increased test knowledge to subsequently make more accurate prediction judgments as the tests continued.

GENERAL DISCUSSION

Robust postdiction superiority effects were found in both experiments, consistent with prior research in metacomprehension (e.g., Glenberg & Epstein, 1985; Glenberg et al., 1987; Maki, Foley, Kajer, Thompson, & Willert, 1990; Maki & Serra, 1992). After reading brief texts, participants were less accurate at predicting their own performance on a given set of four test questions than they were at postdicting after answering the questions. The metacognitive and criterion performance results found in the present study, as well as the materials and methods used, correspond closely to those of other studies of metacomprehension, such as Maki and Serra's (1992) and Weaver and Bryant's (1995). Therefore, the present findings appear to be generalizable to other published studies of metacomprehension accuracy.

What caused the postdiction superiority effect in the present experiments? Retrieval hypotheses state simply that participants remember, while postdicting, how well they answered questions on the previous question set. The robust postdiction superiority effect found in the present experiments is certainly consistent with this class of hypotheses. Furthermore, the finding that neither prediction nor postdiction accuracy improved across question sets or tests is consistent with retrieval hypotheses. Thus, no evidence generated by the present experiments contradicts the class of retrieval hypotheses as an explanation of our observed postdiction superiority effects.

The same claim cannot be made for the test knowledge hypotheses. These hypotheses state that something is learned about the nature of the tests one is to take, and that test knowledge should increase as one takes more tests. For example, as participants take more tests, they might learn more about the difficulty level of the ques-

Table 6
Mean Recognition Performance as a Function of Text Position and Text Selection in Experiment 2

Performance							
Text 1		Text 2		Text 3		All Texts	
M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Text Position							
.500	.021	.517	.024	.478	.039	.500	.016
Text Selection							
.346	.020	.537	.024	.609	.025	.500	.016

Note—Text 1, "The Martian Atmosphere"; Text 2, "Symbiosis"; Text 3, "Euripides."

tions, the specificity of the questions, the deceptiveness of the lures on the test, and so on. In the present experiments, participants had no test knowledge before the first test, and therefore they were expected to produce the postdiction superiority effect seen on the first test. The increasing test knowledge gained on every test, however, should have improved prediction accuracy across tests. The results from both experiments are inconsistent with this prediction of test knowledge hypotheses. Prediction accuracy did not improve across question sets or across texts in either experiment.

These findings may be relevant to the classroom setting. For example, it is commonly believed that although students may founder on the first exam of the semester, their test scores in a new class may improve once they learn more about the exams themselves. Students might learn something about the type of questions to be asked, the detail of understanding of test material that is needed to answer test questions, or the best way to budget time during the exam. Learning about the exams should not only improve subsequent test scores, but should also improve the ability to assess the grade that might be earned for a particular level of comprehension of test material. With such test knowledge, for example, students preparing for an exam might stop studying when they are confident that their comprehension level will earn them the desired grade on the exam.

Although the present study found no evidence that test knowledge improved metacomprehension accuracy, we cannot conclude that our participants, much less students, in general, do not acquire test knowledge as they take more tests of a certain type. If the participants in the present study did acquire test knowledge with successive tests, however, it did not affect the robust postdiction superiority effects that we continued to observe across successive tests.

REFERENCES

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

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition*, **15**, 84-93.

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

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior & Human Performance*, **20**, 159-183.

Maki, R. H. (1998a). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition*, **26**, 959-964.

Maki, R. H. (1998b). 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: Erlbaum.

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, & Cognition*, **16**, 609-616.

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

Metcalfe, J. (1996). Metacognitive processes. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 381-407). San Diego: Academic Press.

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

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

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.

NOTES

1. The term *metamemory accuracy* in this study refers to micropredictive accuracy or discrimination that reflects a participant's ability to know which questions he or she will get correct or incorrect (Schwartz & Metcalfe, 1994). Metcalfe (1996) has referred to the correspondence between participants' ranking of items and later performance on those items as *micrometacognition.* This correspondence, which reflects discrimination among individual items, is normally measured by the nonparametric Goodman–Kruskal gamma correlation, which Nelson (1984) has argued is the appropriate statistic for these types of data. It is important to note that discrimination is independent of overall recognition performance. If one wishes to measure metacognitive accuracy overall without regard to individual items, macroprediction or calibration is the appropriate measure (Schwartz & Metcalfe, 1994). We did, in fact, analyze the data in terms of both discrimination and calibration, which solves a potential problem that arises from the use of gamma correlations; a reduction in sample size due to occasional lack of variability in metacognitive judgments (see note 2). Using the calibration formula from Lichtenstein and Fischhoff (1977), we calculated calibration scores by participant. Analyses of these calibration scores showed the same pattern of results as did the discrimination analysis; there were no significant effects of either question set position or text position on calibration scores.

The failure to find a significant effect of question set or text position on calibration scores was not due to lack of power. The effect size of our overall postdiction superiority effect in Experiment 1 was .26 (a large effect), and in Experiment 2 it was .09 (medium to large effect). These effect sizes are similar to the postdiction superiority effect found by Maki and Serra (1992, Experiment 1), which was .09. To detect an effect size of .09 in our experiments, our power was greater than .90. We chose to report gamma correlations (i.e., discrimination) in the present study to maintain consistency with prior research on metacomprehension accuracy (e.g., Maki, 1998a; Maki & Serra, 1992; Weaver & Bryant, 1995).

2. Individual repeated measures ANOVAs were conducted rather than a full factorial ANOVA because of the large number of missing gamma correlations. These missing values arose because on a number of occasions, participants either made constant metacognitive judgments across the question sets or texts or exhibited constant criterion performance across the question sets or texts. The use of a full factorial ANOVA would have severely restricted the number of observations on which to base our analysis.

3. The levels of performance in the present study were similar to those observed by Weaver and Bryant (1995), who used the same materials. Weaver and Bryant found that for narrative texts, average recognition performance was .64, and for expository texts, it was .41. We found narrative text performance of .72 and expository text performance of .50.

(Manuscript received October 18, 1999;
revision accepted for publication July 13, 2000.)