

## Influences of metamemory on performance predictions for text

Katherine A. Rawson

*University of Colorado at Boulder, Boulder, USA*

John Dunlosky and Susan L. McDonald

*The University of North Carolina at Greensboro, Greensboro, USA*

When predicting future performance on tests over text material, do individuals estimate retention in addition to assessing comprehension? In Experiment 1, participants either rated their comprehension or predicted performance for each text, with lower ratings indicating lower confidence either in comprehension or in eventual performance. Judgement magnitude was significantly lower for performance predictions than for comprehension ratings, suggesting that predictions were based partly on retention estimates. In Experiment 2, predictions varied with anticipated test delay (15 min or 2 weeks) whereas comprehension ratings did not, providing further evidence that individuals estimate retention when predicting performance. Analyses of individual differences suggest that both good and poor performers incorporate retention estimates when predicting performance, but better performers do so in a more discriminative manner. Implications for theory of metacognitive judgements are discussed.

Metacomprehension, or the assessment of one's own comprehension, is a central component of self-regulated comprehension (Hacker, 1998). Consider a student who is studying text material, such as textbook chapters and class notes, in preparation for upcoming examinations. To prepare efficiently for her exams, she should spend less time studying material that she has already learned well so that she may spend more time studying material that she has not learned well. Her ability to regulate her studying effectively will depend in part upon the degree to which she can accurately assess how well she has learned the various text materials. For instance, if she judges that she has learned a particular text adequately when in fact she has not learned the material well enough, she is likely to stop studying that text prematurely. Or, she may judge that she has learned a particular text inadequately when she has learned the material well enough. In this case, she may unnecessarily spend more time studying that text, leaving less time to study material that is less well learned. As this example illustrates, the

---

Requests for reprints should be sent to Dr John Dunlosky, PO Box 26164, Psychology Department, UNCG, Greensboro, NC 27402-6164, USA. Email: [dunlosky@uncg.edu](mailto:dunlosky@uncg.edu)

Many thanks to Ruth Maki for comments concerning an earlier draft of this research.

utility of metacomprehension in self-regulated comprehension depends upon the accuracy of metacomprehension judgements.

Researchers exploring the accuracy of metacomprehension judgements have typically utilized an experimental analogue to the real-world example just given. In these studies, participants read several short texts and then predict for each text how well they will do on a subsequent test over the material. Almost invariably, intra-individual correlations between performance predictions and actual test performance have been quite low. For example, Maki (1998) reported a mean gamma correlation of  $+ .27$  across 25 studies conducted in her laboratory. A probabilistic interpretation of this mean gamma (Nelson, 1984) indicates that our hypothetical student, given two texts that she has learned to different degrees, would be only 14% better than chance at selecting the less well-learned text for restudy. Why are individuals seemingly so poor at making these metacomprehension judgements?

Discovering the bases of judgements may provide insight into why the accuracy of metacomprehension judgements are often low. According to current theory, metacognitive judgements are based on various factors, such as characteristics of the to-be-judged material, information retrieved while making the judgement, beliefs about retention, and so on (e.g., Koriat, 1997; Maki, 1998). That is, people presumably do not have direct access to the underlying representation of the text that is constructed during reading, but instead rely on other cues to infer how well the text has been understood. The accuracy of judgements will increase as the underlying bases for the judgements empirically correlate with subsequent criterion performance (e.g., Koriat, 1993). Thus, utilizing a factor that is predictive of subsequent criterion performance may improve accuracy.

In accord with this inference-based framework, the accuracy of metacomprehension judgements is partly a function of how closely the bases of the judgements relate to test performance. Test performance in most metacomprehension research depends not only upon comprehension of the texts, but also upon memory for the texts because the tests are typically delayed. Therefore, the accuracy of metacomprehension judgements may rely upon the degree to which the judgements are based on an estimate of how well the information will be retained over a delay. That is, this component of metamemory—the estimation of one's future memory for text material—may be one determinant of the accuracy of metacomprehension judgements. If so, the failure of individuals to include a metamemory judgement when predicting performance may attenuate accuracy.

Several researchers have acknowledged the potential importance of a metamemory component to performance predictions for text (Glenberg & Epstein, 1985; Maki, 1998; Maki & Berry, 1984). According to Maki (1998), “[m]aking a prediction about performance on a future test after reading text is a complex process. . . . First, a student must judge how well he or she has understood the text and how much learning has occurred. . . . Second, the student *must predict forgetting* [italics added] that might result both from the retention interval and from learning other material” (p. 119).

Despite the noted importance of retention estimates to the accuracy of performance predictions for text, only one investigation has addressed whether participants incorporate a retention estimate into their performance predictions. Maki and Berry (1984) reported that overall judgement magnitudes were lower for participants predicting their performance on a delayed test (24–72 hours) than for participants predicting for a relatively immediate test (after all text material had been read), suggesting that predictions were sensitive to the

differential memory demands on test performance. However, this pattern was evident only for participants with above-median test performance and was reversed for those with below-median performance.

Given that only one study has examined the role of retention estimates in text predictions and that the results were mixed, the present research was designed to address further the following questions: Do individuals estimate retention when making performance predictions? If so, how are retention estimates incorporated into performance predictions?

To investigate these issues, we made two modifications to the method typically used in metacomprehension research. To investigate if performance predictions include a retention estimate, we compared performance predictions to text judgements that do not require a metamemory judgement. Whereas half of our participants made the typical performance prediction for each text, the other participants were prompted simply to rate how well they understood each text (see also, Maki, Foley, Kajer, Thompson, & Willert, 1990, and Maki & Serra, 1992, which are considered in detail in the General Discussion). The latter group was not forewarned of the subsequent test to avoid the possibility that anticipating the test would induce them to consider the memorability of the texts when making their comprehension ratings. If performance predictions include a retention estimate (i.e., if people anticipate some forgetting), then differences between the two judgement groups will emerge in the magnitude of the judgements. Specifically, performance predictions will be lower than the comprehension ratings that do not require a retention estimate.

We also introduced a test manipulation to provide evidence for how metamemory influences performance predictions. Half of the participants in each judgement group received the equivalent of an open-book test, where each test question was accompanied by its corresponding text. The others received the equivalent of a closed-book test, where each question appeared with only the text title, which is the format typically used in metacomprehension research. Whereas performance on a closed-book test depends upon both memory and comprehension, an open-book test reduces memory demands on performance (Rubenstein, Kender, & Mace, 1988). Comparisons between these groups will allow inferences to be made about how retention estimates are incorporated into performance predictions.

One possibility is that individuals incorporate retention estimates into performance predictions by recognizing that, in general, text material is harder to remember after a delay; thus, they will lower their judgements for all texts. Another possibility is that individuals recognize that particular texts will be less memorable than others, and thus they will differentially adjust the judgement for each text. For example, even though an individual assesses that he understands one text well when reading it, he assigns a value of 20 (out of 100) because he does not expect that he will be able to remember important parts of the text later. For another text, however, he assesses that he does not understand it completely but assigns a value of 60 because he expects to remember what he does understand. In this case, the retention estimates have changed the relationship between the judgements, such that the first text is rated lower than the second text even though the first is understood better at the time of study. If individuals incorporate retention estimates in this discriminative fashion, accuracy will be better for predictions of performance on a closed-book test (in which variable retention of the texts will differentially influence performance) than on an open-book test.

Other comparisons between groups will also be informative with respect to the question of how retention estimates influence performance predictions, which we discuss later. However,

note that the rationale of these analyses relies on the assumption that individuals incorporate retention estimates into performance predictions. Thus, although analyses of accuracy may be informative, our primary interest was in evaluating whether individuals incorporate retention estimates into performance predictions by examining differences in the magnitude of judgements across judgement groups. Finally, as Maki and Berry (1984) found some evidence for individual differences in the extent to which retention estimates influence performance predictions, we also explored the relation between performance level and metacomprehension.

## EXPERIMENT 1

### Method

#### *Participants and design*

A total of 160 undergraduates who were enrolled at the University of North Carolina at Greensboro participated to partially satisfy a course requirement in Introductory Psychology. A total of 40 participants were randomly assigned to each of four experimental groups, formed by the factorial combination of two between-subjects variables: kind of judgement (comprehension rating or performance prediction) and kind of test (open book or closed book).

#### *Materials and apparatus*

Experimental texts were constructed by extracting two sentences from larger expository texts, such that the two sentences together expressed the central principle of the parent text. The parent texts included some used by Glenberg and Epstein (1985), and texts adapted from passages in Graduate Record Examination (GRE) preparation manuals.<sup>1</sup> We developed 13 of these two-sentence experimental texts (including 1 sample text and 12 critical texts), which ranged from 33 to 55 words in length ( $M = 45$ ).<sup>2</sup> Each was assigned an appropriate title.

Two inference verification questions were developed for each experimental text, and each question had both a true and a false version (sample test questions along with corresponding texts appear in the Appendix).<sup>3</sup> The version of each inference to be presented was selected such that each participant received 12 true questions and 12 false questions, and each inference version was used an equal number of times across participants in each group. Order of presentation of inference questions for each participant was determined in the following manner: One test item corresponding to each of the first four texts studied was presented in random order, then one test item corresponding to each of the second four texts was presented in random order, and then one test item corresponding to each of the last four texts was

---

<sup>1</sup>The Graduate Record Examination (GRE) is a standardized test taken by undergraduates prior to admission to graduate school. The GRE includes verbal, quantitative, and analytical components; our materials were taken from short practice texts and test questions for the verbal component.

<sup>2</sup>The texts used in the present research are shorter than those typically used in research on metacomprehension. The overall length of text that experimenters can reasonably ask participants to learn is a methodological constraint. Shorter texts permit the use of more texts, which is preferable for correlational measures. Additionally, Kintsch (1998, chap. 6) argues that the processes operating on shorter experimental texts of this kind are not fundamentally different from those operating on longer texts.

<sup>3</sup>Use of an open-book test required test questions that were not solely memory based, to prevent ceiling-level performance for this group. Inference-based questions have been used in previous research (e.g., Glenberg, Sanocki, Epstein, & Morris, 1987), and as they require both comprehension and memory for the text material, they were most suitable for our purposes.

presented in random order. The second test item corresponding to each text was presented in the same manner. Macintosh computers presented instructions and experimental materials and recorded all data.

### *Procedure*

All participants were instructed that they would be reading sentence pairs and that they should do their best to understand the central principle expressed in each pair. Participants were shown the sample text, and they were prompted to practise making the judgement appropriate to their condition. Participants making performance predictions were told that they would later complete an inference verification test for each text, and they were shown a sample test question. Those who would later receive an open-book test were not told beforehand that the text would accompany each question. Participants making comprehension ratings were not forewarned of the test. Instead, they were told that we were “developing materials for a reading comprehension study and needed to know if they were at an appropriate level of difficulty.” Each individual was explicitly instructed to rate his or her own understanding of each text.<sup>4</sup>

Each critical text was then presented individually along with its title for self-paced study. The order of presentation was randomized anew for each participant. Each text remained on the screen until the participant pressed a key to advance. Immediately following the offset of a text, the participant was prompted to make a judgement for that text. Performance predictions were prompted with the title of the text and the query, “How confident are you that you will be able to correctly evaluate an inference question based on this text in about 15 minutes? 0 = definitely won’t be able, 20 = 20% sure you will be able, 40 = 40% sure . . . , 60 . . . , 80 . . . , 100 = definitely will be able.” Comprehension ratings were prompted with the title of the text and the query, “How well do you understand the information presented in this text? 0 = definitely don’t understand, 20 = 20% sure you understand, 40 = 40% sure, 60 . . . , 80 . . . , 100 = definitely understand.”

After all texts had been studied and judged, half of the participants in each judgement group received a closed-book test, and the other half of each group received an open-book test. For the closed-book test, each inference verification question appeared individually on the screen, along with the title of the corresponding text. For the open-book test, each inference verification question appeared individually, along with both the title and the corresponding text. For both test groups, participants were instructed to decide whether each inference was “true” or “false” according to the information contained in the text. Participants indicated their responses by clicking on either the “TRUE” or the “FALSE” button appearing on the screen along with each question. After answering each question, participants were prompted to make a postdiction (i.e., indicate their confidence in the accuracy of their response) with the query, “For the inference question entitled \_\_\_\_\_, how confident are you that you answered correctly? 0 = don’t know if you were correct, 20 = 20% sure you were correct, 40 = 40% sure . . . , 60 . . . , 80 . . . , 100 = definitely was correct.”

## Results and discussion

*Test performance.* For each participant, the mean test score across the 12 texts was computed. Across individual means, median performance for the closed-book test was .61 for participants who predicted performance and .58 for those who rated comprehension. Median performance for the open book test was .67 for participants who predicted performance and .67 for those who rated comprehension. A 2 (kind of judgement)  $\times$  2 (kind of test) analysis of

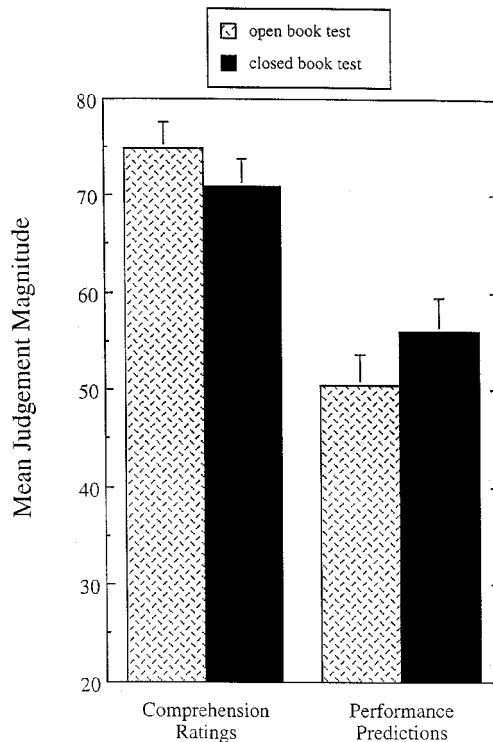
---

<sup>4</sup>Note that if these incidental instructions were not effective (i.e., participants making comprehension ratings nonetheless suspected a test), then the differences in magnitude would be attenuated between comprehension ratings and performance predictions in a way that would be inconsistent with our focal hypothesis.

variance (ANOVA) indicated that the main effects and the interaction were not significant, all  $F$ 's  $< 1.64$ . Although scores for each group were low,  $M$ s  $> 0.58$ ,  $SEM$ s = 0.02, they were significantly greater than chance and thus allowed interpretation of the analyses of judgement accuracy.

The finding that performance was not significantly better on the open-book test than on the closed-book test may have been due to the nature of our inference test questions, which depended not only on memory but also on comprehension of the text information. A larger difference between performances for open-book and closed-book tests may have manifested for tests that relied more exclusively on memory (e.g., verbatim recognition questions). Nonetheless, the trend suggests that performance on closed-book tests was somewhat more dependent on memory for the material than that on the open-book test.

*Magnitudes of the judgements.* For each participant, the mean judgement across the 12 texts was calculated. For each group, the mean across individual means is shown in Figure 1. A 2 (kind of judgement)  $\times$  2 (kind of test) ANOVA revealed a significant effect of kind of judgement,  $F(1, 156) = 41.10$ , whereas the effect of test and the interaction were not significant,  $F < 1$  and  $F(1, 156) = 2.38$ , respectively. Thus, performance predictions were significantly less than comprehension ratings, which confirms the hypothesis that performance predictions include retention estimates.



**Figure 1.** For Experiment 1, mean judgement magnitudes across individual means are shown as a function of kind of judgement and kind of test. The standard error of the mean is presented for each condition.

*Magnitudes of the judgements as a function of performance level.* As previously discussed, Maki and Berry (1984) found that individuals with above-median test performance made lower predictions for a delayed test than for an immediate test, whereas individuals with below-median performance made higher predictions for the delayed test. Consequently, we conducted analyses to investigate whether the differences in judgement magnitudes between performance predictions and comprehension ratings were localized to a subset of our participants. As performance level did not significantly differ by group, a median split based on performance across all groups was used ( $Mdn = .63$ ). Participants performing at the median were included in the above-median group, as in Maki and Berry. This median split for participants who made performance predictions resulted in 47 participants above the median and 33 below the median. The median split for participants who made comprehension ratings resulted in 41 participants above the median and 39 below the median. Also, because the kind of test did not influence the overall magnitude of judgements, we collapsed across this variable for the present analyses. For each group, the mean across individual mean judgements is reported in Table 1.

A 2 (kind of judgement)  $\times$  2 (performance level) ANOVA revealed significant main effects for kind of judgement,  $F(1, 156) = 46.35$ , and performance level,  $F(1, 156) = 12.29$ . However, the interaction was not significant,  $F < 1$ . Thus, above-median performers tended to make higher judgements than did below-median performers. More important, comprehension ratings were greater than performance predictions both for above-median performers and for below-median performers. These results do not conceptually replicate those from Maki and Berry (1984), an issue addressed further in the General Discussion.<sup>5</sup>

TABLE 1  
Mean judgement magnitudes by performance level and by judgement group for Experiments 1 and 2

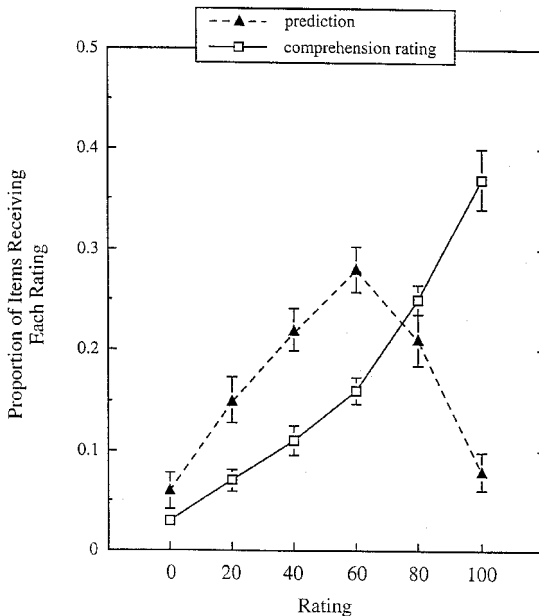
<i>Experiment</i>	<i>Performance level</i>	<i>Test delay</i>	<i>Above-median</i>		<i>Below-median</i>	
			<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
1	Performance predictions		57	2.7	47	4.2
	Comprehension ratings		78	2.5	67	2.8
2	Performance predictions	15 min	59	4.2	45	3.9
		2 weeks	48	4.1	40	3.9
	Comprehension ratings	15 min	65	4.4	49	4.8
		2 weeks	61	4.3	50	6.3

<sup>5</sup>We used median split analyses to examine judgement magnitudes as a function of performance to facilitate comparison with results reported by Maki and Berry (1984). However, dichotomizing a single, continuous predictor variable reduces statistical power, relative to a regression analysis that involves the continuous nature of the variable (see Maxwell & Delaney, 1993, for relevant discussion, as well as explanation of circumstances under which median split analyses can spuriously inflate statistical significance). Thus, to allow for the possibility that reduced statistical power did not allow us to detect the interaction evident in Maki and Berry's results, we also analysed judgement magnitudes using an analysis of covariance (ANCOVA). The interaction between kind of judgement and performance was not significant in this analysis either.

*Proportion of texts receiving each judgement rating.* Examination of the distributions of the judgements affords a more detailed characterization of how retention estimates influenced performance predictions. For each participant, the proportion of texts receiving each of the judgement values was calculated. Because the kind of test did not influence the magnitude of the judgements or the distributions, we collapsed these analyses across the kind of test. The mean proportions across participants are presented in Figure 2.

The use of the rating scale was qualitatively different for individuals making predictions than for those making comprehension ratings. Namely, predictions showed an inverted-U-shaped pattern, whereas the proportion of items receiving each comprehension rating increased monotonically from the lowest to highest rating. Our proposal is that these differences provide converging evidence that performance predictions include a retention estimate. Individuals apparently believe that they have excellent comprehension of the majority of texts immediately after reading (based on the distribution of comprehension ratings). However, when asked to predict later performance, they are less confident that their understanding of the material will be retained.

The present pattern of distribution is also consistent with findings based on judgements of learning made for paired associates. Across two experiments, Dunlosky and Nelson (1994) had participants study paired associate (e.g., dog-spoon) and judge either (1) the likelihood that they would later retrieve each item during a future test (Experiment 1) or (2) how well each item had been learned (Experiment 2). When individuals predicted future performance, the distribution of ratings showed an inverted-U-shaped pattern (cf., predictions in Figure 2). By contrast, when individuals assessed how well each item had been learned, the proportion of



**Figure 2.** For Experiment 1, proportions of text items receiving each value on the rating scale are shown as a function of kind of judgement. The standard error of the mean is presented for each value.

items receiving each judgement increased monotonically from lowest to highest rating (cf., comprehension ratings in Figure 2). Although a variety of factors other than the kind of rating differed across the two experiments, Dunlosky and Nelson speculated that for the judgements of learning made in their Experiment 2, "the subjects may not have incorporated their theory of retention" (p. 560), which is analogous to the interpretation of the present findings.

*Judgement accuracy.* Judgement accuracy was operationalized as the gamma correlation between an individual's judgements and test scores across texts. The mean across individual gammas for each group was computed. For participants making performance predictions, mean gammas were .40 ( $SEM = 0.07$ ) for the closed-book test and .19 ( $SEM = 0.07$ ) for the open-book test. For participants making comprehension ratings, mean gammas were .18 ( $SEM = 0.08$ ) for the closed book test and .23 ( $SEM = 0.06$ ) for the open book test. A 2 (kind of judgement)  $\times$  2 (kind of test) ANOVA indicated that whereas neither the main effect of judgement nor that of test was significant,  $F(2, 153) = 1.61$ , and  $F(1, 153) = 1.30$ , respectively, the interaction approached statistical reliability,  $F(1, 153) = 3.29$ ,  $p = .07$ .

Comparison of the accuracy of performance predictions for participants who received a closed-book test with the accuracy for those who received an open-book test was of particular interest for evaluating how retention estimates are incorporated into performance predictions. Follow-up tests to the interaction revealed that performance predictions were significantly more accurate for participants who received a closed-book test than for those who received an open-book test,  $t(76) = 2.12$ . This finding suggests that participants incorporated retention estimates by differentially adjusting their predictions to allow for variable forgetting across texts. If participants had incorporated retention estimates by lowering all predictions uniformly, each prediction would have maintained the same relationship to the others as they would have if retention estimates had not been incorporated. If so, comprehension ratings and performance predictions would be equally predictive of performance on a closed-book test. However, performance predictions were significantly more predictive of performance on the closed-book test than were comprehension ratings,  $t(75) = 2.08$ . This result replicates a finding reported by Maki and Serra (1992). They had participants rank several texts based either on the anticipated performance level for each or on the ease of comprehending each. Ranks based on anticipated performance were more predictive of actual performance than were ranks based on comprehension ease. Taken together, the present and past findings converge on the conclusion that retention estimates are incorporated into performance predictions discriminatively.

*Relation between metacomprehension accuracy and performance level.* Maki and Berry (1984) reported that across individuals, judgement accuracy increased as performance on a delayed test increased. However, accuracy did not vary as a function of performance on an immediate test. They concluded that better performers were more skilled at factoring forgetting into their performance predictions than were poorer performers, whereas the two groups were equally skilled at assessing comprehension. Although the pattern of judgement magnitudes in our Experiment 1 indicated that both better and poorer performers were sensitive to memory demands on test performance (Table 1), better and poorer performers may incorporate retention estimates into their performance predictions differently. For instance, better performers may differentially use retention estimates in their performance predictions to dis-

criminate between texts. If so, the accuracy of performance predictions would be expected to improve as performance improved, although only for the closed-book (memory-dependent) test. With respect to comprehension ratings, if both groups are equally skilled at assessing comprehension (as suggested by Maki & Berry), judgement accuracy will not differ as a function of performance for those rating comprehension.

To explore this possibility, we calculated the Pearson product-moment correlation between test performance and judgement accuracy for each of the four experimental groups (as in Maki & Berry, 1984). For participants predicting performance on the closed-book test, accuracy was significantly correlated with performance level,  $r = .45$ . By contrast, this correlation was not significant for participants predicting performance on the open-book test,  $r = .12$ , or for those who rated comprehension either on the open-book test,  $r = .16$ , or on the closed-book test,  $r = .01$ . Thus, judgement accuracy showed the largest relation to performance when participants were asked to predict their performance on a closed-book test. These outcomes suggest that better performers incorporate theory of retention into predictions in a fashion that discriminates between performance across texts, whereas poorer performers do this to a lesser degree. The overall pattern of accuracy as a function of performance level also supports Maki and Berry's conclusions.

*Magnitudes of the postdictions.* For each participant, the mean postdiction across the 24 test questions was calculated. Across individual means, mean postdiction for the closed-book questions was 60 ( $SEM = 3$ ) for participants who predicted performance and 63 ( $SEM = 2$ ) for those who rated comprehension. Mean postdiction for the open-book questions was 60 ( $SEM = 3$ ) for participants who predicted performance and 71 ( $SEM = 3$ ) for those who rated comprehension. A 2 (kind of judgement)  $\times$  2 (kind of test) ANOVA revealed a significant main effect of kind of judgement,  $F(1, 156) = 6.41$ , whereas neither the main effect of kind of test nor the interaction was significant,  $F_s < 1.90$ . Thus, postdictions were greater for individuals who had rated comprehension than for those who had predicted performance. This finding suggests that some of the difference between the magnitudes of predictions and comprehension ratings (shown in Figure 1) may be attributable to a general tendency for individuals who rated comprehension to make higher judgements. Even so, the absolute difference in postdiction magnitudes for the two groups was considerably smaller than the difference in pre-test judgement magnitudes (7 vs. 20 percentage points, respectively), as confirmed by a significant interaction when judgement magnitudes were submitted to a 2 (kind of predictive judgement)  $\times$  2 (predictive versus postdictive judgement) ANOVA,  $F(1, 158) = 34.87$ ,  $MSe = 87.24$ . This outcome suggests that the general tendency for individuals who rated comprehension to make higher judgements cannot account entirely for the differences between comprehension ratings and performance predictions.

## EXPERIMENT 2

Comparison of the magnitudes of performance predictions and comprehension ratings in Experiment 1 suggested that individuals incorporate retention estimates into predictions. Although comparison of different kinds of judgement has been used previously to infer the bases of various metacognitive judgements, factors other than those of experimental interest can also yield differences in magnitudes. Our conclusion that performance predictions

incorporate retention estimates, which was based on differences between magnitudes for predictions and comprehension ratings, would be bolstered by further dissociation of the two judgements.<sup>6</sup> In particular, demonstrating that the two kinds of judgement differ along a dimension that is theoretically related to estimating retention, such as the length of delay between judgement and test, would provide more definitive support for the conclusion.

In Experiment 2, participants either made performance predictions or rated comprehension for each text. Most important, all participants were informed that they would be tested on half of the texts after a 15-min delay, whereas the remaining texts would be tested after 2 weeks. If performance predictions include retention estimates, then judgement magnitudes will be lower when a 2-week delay is anticipated than when a 15-min delay is anticipated. In contrast, magnitudes for ratings of comprehension were not expected to differ as a function of anticipated delay.

## Method

### *Participants and design*

A total of 80 undergraduates who were enrolled at the University of North Carolina at Greensboro participated to partially satisfy a course requirement in Introductory Psychology. A total of 40 participants were randomly assigned to each of two experimental groups, as defined by kind of judgement (comprehension rating or performance prediction). Anticipated test delay (15 min or 2 weeks) was a within-participant manipulation.

### *Materials and apparatus*

Materials from Experiment 1 were used. Macintosh computers presented instructions and experimental materials and recorded all data.

### *Procedure*

All participants were instructed that they would be reading sentence pairs and that they should do their best to understand the central principle expressed in each pair. Participants were shown the sample text, and they were prompted to practise making the judgement appropriate to their group. All participants were told that they would later complete an inference verification test for each text. They were also told that half of the texts would be tested after a 15-min delay, whereas the other half of the texts would be tested after a 2-week delay.

Each critical text was then presented individually along with its title for self-paced study. The order of presentation was randomized anew for each participant. Each text remained on the screen until the participant pressed a key to advance. Immediately following the offset of a text, the participant was prompted to make the judgement appropriate to their group for that text. Each judgement prompt was accompanied by notification of whether the text would be tested after 15 min or after 2 weeks. Because our critical hypothesis concerned the influence of anticipated delay on the two kinds of judgement and not on judgement accuracy, we did not administer test questions at the 2-week delay. Instead, immediately after all texts had been studied and judged, all test questions were administered in the closed-book format. Participants made a postdiction after answering each question.

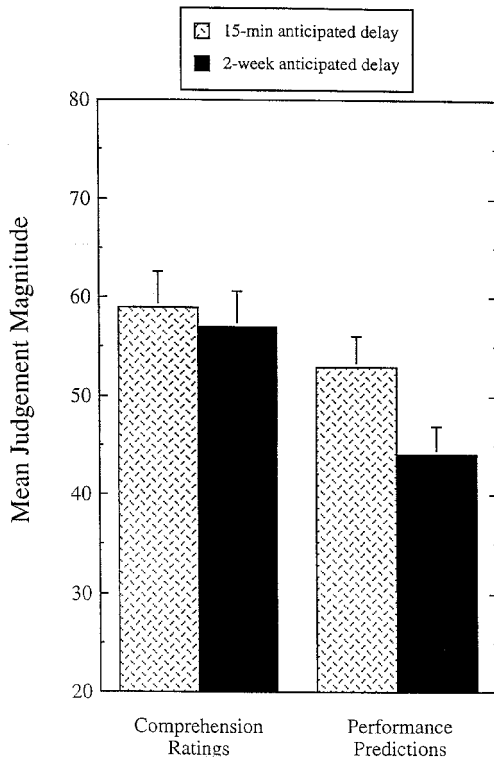
---

<sup>6</sup>We thank Asher Koriat for this idea.

## Results and discussion

*Test performance.* For each participant, the mean test score across the 12 texts was computed. Across individual means, median performance was .58 for participants who predicted performance and .63 for those who rated comprehension. Performance for the two groups did not significantly differ,  $t(78) = 1.25$ . Again, overall performance was relatively low,  $M_s > 0.56$ ,  $SEMs = 0.02$ , but significantly greater than chance.

*Magnitudes of the judgements.* For each participant, the mean judgement across the 12 texts was calculated. For each group, the mean across individual means is shown in Figure 3. A  $2$  (kind of judgement)  $\times 2$  (anticipated delay) ANOVA revealed significant main effects of kind of judgement,  $F(1, 78) = 4.72$ , and anticipated delay interval,  $F(1, 78) = 13.8$ . Importantly, these main effects were qualified by a significant interaction,  $F(1, 78) = 4.27$ . Follow-up tests indicated that performance predictions were lower when a 2-week delay was anticipated than when a 15-min delay was anticipated,  $t(39) = 4.01$ , whereas comprehension ratings did not differ significantly with anticipated delay,  $t(39) = 1.18$ . This outcome provides further support for the hypothesis that performance predictions include retention estimates.



**Figure 3.** For Experiment 2, mean judgement magnitudes across individual means are shown as a function of kind of judgement and kind of test. The standard error of the mean is presented for each condition.

*Magnitudes of the judgements as a function of performance level.* As in Experiment 1, we used a median split on test performance to analyse whether above-median and below-median performers differentially utilized retention estimates when predicting performance. As performance level did not significantly differ by judgement group, a median split based on performance across all participants was used ( $Mdn = .58$ ). Again, participants performing at the median were included in the above-median group. The median split for participants who made performance predictions resulted in 22 participants above the median and 18 below the median. The median split for participants who made comprehension ratings resulted in 27 participants above and 13 below the median. For each group, the mean across individual mean judgements is reported in Table 1.

A 2 (kind of judgement)  $\times$  2 (performance level)  $\times$  2 (anticipated delay) ANOVA revealed a marginal main effect of kind of judgement,  $F(1, 76) = 3.09, p = .08$ , and significant main effects of performance level,  $F(1, 76) = 7.68$ , and anticipated delay,  $F(1, 76) = 10.46$ . The Judgement  $\times$  Delay interaction was significant,  $F(1, 76) = 4.74$ , which is redundant with the previous analysis, indicating that performance predictions differed with anticipated delay but comprehension ratings did not. The Performance Level  $\times$  Delay interaction approached statistical significance,  $F(1, 76) = 3.81, p = .06$ . Finally, neither the Judgement  $\times$  Performance Level interaction nor the three-way interaction was significant,  $F_s < 1$ .

Of potential interest is the interaction of performance level with anticipated delay. Follow-up tests indicated that above-median performers made lower judgements when anticipating a 2-week delay than when anticipating a 15-min delay,  $t(48) = 4.57$ , whereas judgements for below-median performers did not differ with anticipated delay,  $t(30) = 0.83$ . This interaction is consistent with findings from Experiment 1 and with those reported by Maki and Berry (1984), which suggest that above-median performers incorporate retention estimates in a discriminative fashion whereas below-median performers do so to a lesser extent. However, the present results must be interpreted with some caution because even for below-median performers, predictions were greater for 15-min items than for 2-week items (although not significantly so), and because the potential interaction of performance level with anticipated delay is scale dependent (Loftus, 1978).

*Proportion of texts receiving each judgement rating.* As in Experiment 1, the proportion of texts receiving each of the judgement values was calculated for each participant. Mean proportions across participants are presented in Figure 4.

Although the distributions for comprehension ratings do not show the same pattern as that in Experiment 1, a key difference in scale use for comprehension ratings vs. performance predictions is still evident. Namely, people more often used the highest rating of 100 when making comprehension judgements than when making predictions: When the proportion of texts assigned the value of 100 was submitted to a 2 (kind of judgement)  $\times$  2 (anticipated delay) ANOVA, a significant main effect of kind of judgement was obtained,  $F(1, 75) = 6.59, MSE = 0.05, p < .05$ . The main effect of anticipated delay approached statistical significance,  $F(1, 75) = 3.83, MSE = 0.01, p = .05$ , and the interaction was not significant,  $F < 1$ .

*Judgement accuracy.* For each individual, a gamma correlation between judgements and test scores across texts was calculated. The mean across individual gammas was .13 ( $SEM = 0.07$ ) for participants who made performance predictions and .12 ( $SEM = 0.07$ ) for those who

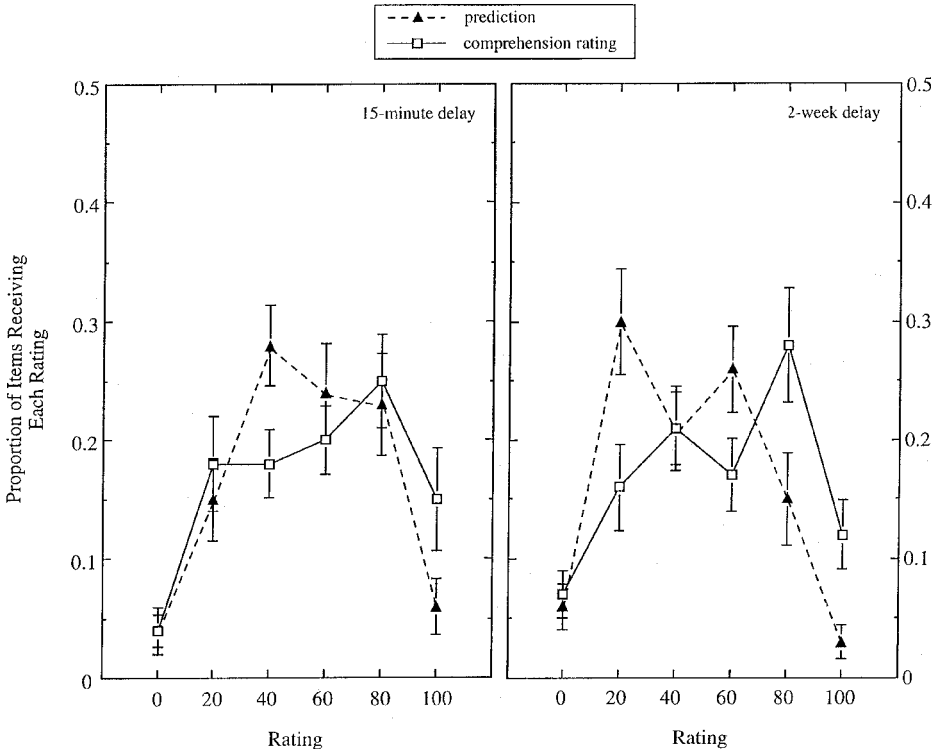


Figure 4. For Experiment 2, proportions of text items receiving each value on the rating scale are shown as a function of kind of judgement and anticipated delay between reading and the test. The standard error of the mean is presented for each value.

made comprehension ratings. These means did not differ,  $t(75) = 0.10$ . The low level of accuracy in both groups was not surprising, given the methodology used in this experiment. For instance, comprehension ratings presumably did not include retention estimates but yet were compared with test performance that was partly memory dependent. Predictions included retention estimates but differentially so, depending on whether a relatively short or long delay was anticipated. However, judgements were compared with performance from tests that were all administered after a short delay. Thus, perhaps judgements made in anticipation of a 15-min delay would be more accurate than those made in anticipation of a 2-week delay, as the former would more closely reflect the actual delay interval. Consistent with this possibility, predictions for 15-min items tended to be more accurate (mean gamma = .20,  $SEM = 0.11$ ) than those for 2-week items (mean gamma = .01,  $SEM = 0.11$ ), although not significantly so.

*Relation between metacomprehension accuracy and performance level.* For purpose of completeness, we again report Pearson correlations between test performance and judgement accuracy across individuals in each group. For participants predicting performance, accuracy was significantly correlated with performance level,  $r = .33$ . As in Experiment 1, this correlation was not significant for participants who rated comprehension,  $r = -.10$ .

*Magnitudes of the postdictions.* The mean postdiction for the 15-min items was 54 ( $SEM = 3$ ) for participants who predicted performance and 51 ( $SEM = 4$ ) for those who rated comprehension. Mean postdiction for the 2-week items was 54 ( $SEM = 3$ ) for participants who predicted performance and 52 ( $SEM = 4$ ) for those who rated comprehension. Neither of the main effects nor the interaction was significant,  $F_s < 1$ . These outcomes indicate that the differences between comprehension ratings and performance predictions examined earlier were not due to a general tendency for individuals who rated comprehension to make higher judgements.

## GENERAL DISCUSSION

### Bases of performance predictions: Retention estimates vs. knowledge of test

Evidence from two experiments converges on the conclusion that performance predictions include retention estimates. In Experiment 1, judgement magnitudes were significantly lower for participants who predicted performance than for those who rated comprehension, suggesting that participants anticipated some forgetting when predicting performance. This conclusion is also supported by evidence from previous research. Maki et al. (1990) investigated how judgement accuracy was influenced by ease of processing, which they manipulated by presenting intact texts vs. texts in which the letters of some words were deleted. In one experiment, Maki et al. had one group of participants rate ease of comprehension and another group predict test performance. Similar to the results presented here, the magnitude of ease-of-comprehension ratings tended to be greater than the magnitude of test predictions, although the statistical significance of this trend was not reported.

In addition to the differences in judgement magnitudes, the distributions of judgements in the present research also revealed qualitative differences between how people predict performance vs. rate comprehension. In Experiment 1, the value of 100 was the most frequently assigned comprehension rating; individuals who predicted performance assigned this value less often than almost all other values, resulting in a crossover interaction (Figure 2). Although the qualitative shift in use of the two rating scales is not evident in Experiment 2, the same crossover interaction was apparent across the middle-to-highest values of the ratings (Figure 4). These outcomes underscore the influence of retention estimates on text predictions.

Although individuals apparently incorporate a retention estimate into their performance predictions, the differences in judgement magnitudes and distributions could also have been due to people's knowledge about the nature of the test. That is, knowledge of a test may also influence an individual's predictions of test performance. Maki (1998) has noted that in addition to assessing comprehension and estimating retention, "predictions will depend on correct knowledge about the nature of the test, including the types of questions and their level of difficulty" (p. 119). In Experiment 1, participants who rated comprehension were not aware that they would later be tested, whereas those who predicted performance anticipated a test and read a sample question similar to those that they would later receive. Thus, the differences that emerged between the judgement groups may have been a result of differential consideration of the nature of the test rather than differential estimation of retention. For instance, the lower

judgements of those who predicted performance may have been due to the perceived difficulty of the sample inference question that was provided in the pre-experimental instructions.

Support for the contribution of test expectancy to judgement differences may be derived from comparison of the magnitudes of comprehension ratings for Experiments 1 and 2. Unlike Experiment 1, participants who rated comprehension in Experiment 2 were not only forewarned of an upcoming test but were continuously reminded of it with each judgement prompt. Not surprisingly, comprehension ratings for Experiment 2 were lower than those for Experiment 1. As the two groups were similar in all other important respects, this difference in magnitudes is most reasonably attributed to an influence of differential knowledge of the test.

Even though test expectancy may affect comprehension ratings, it cannot be the sole basis of differences between the magnitudes of performance predictions and comprehension ratings. Importantly, individuals in both judgement groups in Experiment 2 were provided with the same information about the subsequent test. If differential knowledge of the test had produced the magnitude differences in Experiment 1, we would have expected this difference to disappear in Experiment 2, which was not the case.

Given evidence that performance predictions include retention estimates, our second goal was to examine how retention estimates are incorporated into performance predictions. People's accuracy at predicting performance indirectly bears on this issue. Namely, if people shift all of the judgements downward when estimating retention, the relative accuracy of the judgements will not be affected. In this case, people merely subtract a constant from each judgement to reflect the overall influence of forgetting on performance, which would not influence our correlation-based measure of relative accuracy. By contrast, people may not only shift predictions downward, but their predictions may also discriminate between the differential forgetting of the texts. If people's predictions are discriminative with respect to differential forgetting, the accuracy of predictions will be greater when forgetting is more likely to influence performance differentially across texts (as on a closed-book test) than when it is less likely to influence performance differentially (as on an open-book test). Results from Experiment 1 provided initial evidence that individuals use retention estimates discriminatively. Namely, performance predictions were more accurate at predicting performance on closed-book tests than on open-book tests.

## Individual differences in the use of retention estimates

Our contention that performance predictions include retention estimates is also supported by outcomes from Maki and Berry (1984), in which participants tended to predict that performance would be greater for an immediate test than for a delayed test. However, Maki and Berry found that only above-median performers predicted lower performance for the delayed test than for the immediate test. By contrast, in the present research, both above-median performers and below-median performers appeared to estimate retention, given performance predictions were lower than comprehension ratings in both groups.

A closer comparison of the methods used in the present research with those used in Maki and Berry (1984) provides potential explanations for this apparent inconsistency. Maki and Berry had participants read a lengthy textbook chapter and make judgements for each section within, and they were tested with a multiple-choice test. In the present experiments, participants made judgements for short texts that were independent from one another in content and

were tested with inference verification questions. Although these methodological differences may have differentially influenced the way in which better and poorer performers made their judgements, we suggest that the pattern of findings across the two studies more strongly implicates the delay interval between study and test as the critical factor.

Specifically, Maki and Berry's (1984) "immediate" test was administered after all texts had been read, and hence it is conceptually identical to our closed-book test in Experiment 1. Thus, participants in both experiments made judgements that required estimating retention over a short delay. An important difference between the two experiments is in the comparison groups. In Experiment 1, we compared judgements requiring a retention estimate over a short delay with comprehension ratings, which are retrospective judgements that presumably do not include an estimate of retention. By contrast, Maki and Berry compared judgements requiring a retention estimate over a short delay with judgements requiring a retention estimate over a much longer delay (e.g., up to 72 hours). Thus, whereas our comparison was relevant to whether individuals anticipate that forgetting will constrain performance on a memory-dependent test, Maki and Berry's comparison was relevant to whether individuals anticipated increased forgetting with increased delays.

Based upon these methodological differences, we offer the following interpretation to reconcile the apparent dissimilarities in the outcomes from these two studies: Both above-median and below-median performers recognize that forgetting constrains performance on a memory-dependent test. However, only above-median performers anticipate the extent to which forgetting will differentially influence test performance across longer retention intervals. This interpretation is consistent with outcomes from Experiment 2 in which individuals made ratings when anticipating either a short (15-min) or long (2-week) delay prior to the test. In this case, the anticipated delay had a greater influence on the judgements made by above-median performers than those made by below median performers.

These results suggest that predictions are influenced by an individual's knowledge about forgetting, and hence the method presented here may provide a measure of this knowledge. One application involves describing the development of this knowledge early in life and the subsequent age-related changes that may occur in later life (e.g., Dixon & Hulstsch, 1983; Kreutzer, Leonard, & Flavell, 1975). Consider the relevant literature on cognitive ageing. When predicting performance, relative accuracy of the predictions is age-invariant (Connor, Dunlosky, & Hertzog, 1997). However, such age invariance does not necessitate that older and younger adults have the same knowledge about memory. That is, an older adult may be able to discriminate among well-learned and poorly-learned materials (as measured by relative accuracy), but not have a complete understanding of how certain factors (e.g., different retention intervals or different kinds of distractor tasks) influence forgetting. If so, these factors will influence older adults' predictions less than younger adults' predictions. Note, however, that the previous approaches to evaluating knowledge and those used in present research are critically different. In particular, predictions presumably depend not just on whether an individual has knowledge but also on whether he or she utilizes it in making the judgements (Dunlosky & Hertzog, 2000). The measures currently in use are questionnaires that tap whether an individual has knowledge (Hertzog, Hulstsch, & Dixon, 1989). Accordingly, the previous methods and the present one can provide complementary evidence concerning an individual's knowledge about retention (questionnaires and predictions) and its utilization in predicting performance during learning (predictions).

## Prompts for metacognitive judgements: Changing the question can change the answer

When people predict test performance for text, they base those predictions in part on their estimate of retention. This conclusion has implications for understanding how people make other kinds of metacognitive judgement. Namely, subtle differences in how experimenters prompt people to make metacognitive judgements may have a substantive impact on how people make them (Kelemen, 2000). For instance, to investigate judgements of learning, experimenters may ask people to judge how well they had learned each item or to predict performance of each item on a subsequent test (e.g., Dunlosky & Nelson, 1994). Although both judgements have been considered “judgements of learning,” the present research suggests that when people are asked to predict performance, they do more than simply attempt to judge their learning of an item. That is, individuals not only infer how well an item has been learned but also consider the likelihood of it being retained. This observation is somewhat ironic in that judgements of learning are typically collected by having people predict future memory performance, and hence making them will involve more than “judging one’s learning” per se.

Of course, we are not arguing that researchers should rename terms that have become standard in this research area (e.g., feelings of knowing and judgements of learning) or that they should not have individuals predict future test performance. On the contrary, investigating predictions of test performance is arguably most germane to improving student scholarship, which relies on knowing how well one will do on an upcoming exam. Instead, the present research demonstrates that progress can be made by systematically investigating how various prompts influence specific metacognitive judgements.

In summary, we evaluated whether one factor—retention estimates—influenced people’s predictions of test performance for text material (for a discussion of other factors, see Maki, 1998, and Rawson & Dunlosky, in press). We analysed differences between performance predictions and comprehension ratings on various dependent measures (judgement magnitudes, distributions, and accuracy). Outcomes from these analyses provided converging evidence for the conclusion that metamemory influences judgements of comprehension.

## REFERENCES

- Connor, L.T., Dunlosky, L., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*, 50–71.
- Dixon, R.A., & Hultsch, D.F. (1983). Structure and development of metamemory in adulthood. *Journal of Gerontology, 38*, 682–688.
- Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging, 3*, 462–474.
- Dunlosky, J., & Nelson, T.O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*, 545–565.
- Glenberg, A.M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and 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.
- Hacker, D.J. (1998). Self-regulated comprehension during normal reading. In D.J. Hacker, J. Dunlosky, & A.C. Graesser (Eds.), *Metacognition and educational theory and practice* (pp. 165–191). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

- Hertzog, C., Hulstsch, D.F., & Dixon, R.A. (1989). Evidence for the convergent validity of two self-report metamemory questionnaires. *Developmental Psychology, 25*, 687–700.
- Kelemen, W.L. (2000). Metamemory cues and monitoring accuracy: Judging what you know and what you will know. *Journal of Educational Psychology, 92*, 800–810.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.
- Koriat, A. (1993). How do we know what we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609–639.
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370.
- Kreutzer, M.A., Leonard, C., & Flavell, J.H. (1975). An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development, 40*, 1–57.
- Loftus, G.R. (1978). On interpretation of interactions. *Memory & Cognition, 6*, 312–319.
- Maki, R.H. (1998). Test predictions over text material. In D.J. Hacker, J. Dunlosky, & A.C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–144). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Maki, R.H., & Berry, S.L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 663–679.
- 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., & Serra, M. (1992). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 116–126.
- Maxwell, S.E., & Delaney, H.D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*, 181–190.
- Nelson, T.O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.
- Rawson, K.A., & Dunlosky, J. (in press). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Rubenstein, H., Kender, J.R., & Mace, F.C. (1988). Do tests penalize readers for poor short term memory? *Journal of Reading, 4*–10.

Original manuscript received 17 March 2000

Accepted revision received 25 May 2001

## APPENDIX

### Sample texts and test question

#### Sample text

Title: Enhancing the Apparent Size of Listening Space

The Bose 901 loudspeaker radiates most of its sound not toward the listener but toward the wall in the back of the speaker and toward the sides of the room. Hearing sounds reflected from the walls of a room increases the apparent size of the room.

Title: Viruses and Cancer

Because no evidence exists that human cancer is infectious, many contend that viruses cannot possibly be significant causal agents in the development of human cancer because viruses are infectious agents. However, a virus that never kills its host can mutate to form a new strain that always kills its host.

#### Sample test questions

Words in parentheses indicate those changed to form the “true” and “false” versions of each question.

Title: Enhancing the Apparent Size of Listening Space

1. In a room that has walls covered in sound absorbing drapes, Bose speakers (would not/would) increase the apparent size of the room.

2. Bose speakers (would not/would) be effective at enhancing perceived listening space if used during an outdoor performance.

Title: Viruses and Cancer

1. It is (possible/impossible) that exposure to chemicals that are non-carcinogenic but are known to cause viral mutations could increase the chances of developing cancer anyway.

2. Medicinal drugs that are designed to operate on infectious agents such as mutating viruses (may/may not) also combat non-infectious cancers.

Copyright of Quarterly Journal of Experimental Psychology: Section A is the property of Psychology Press (T&F) and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.