



# What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses

John Dunlosky<sup>a,\*</sup>, Katherine A. Rawson<sup>a</sup>, Erica L. Middleton<sup>b</sup>

<sup>a</sup> Department of Psychology, Kent State University, Kent, OH 44242, USA

<sup>b</sup> University of Illinois at Urbana-Champaign, IL 61820 USA

Received 30 August 2004, revision received 16 January 2005

Available online 2 March 2005

## Abstract

We evaluated two hypotheses—transfer appropriate monitoring (TAM) and the accessibility hypothesis—that explain why the accuracy of metacomprehension judgments is commonly low. In 2 experiments, participants read six expository texts, made global judgments about how well they would perform on a test over each text, and made term-specific judgments for predicting the recall of definitions embedded in each text. Criterion tests involved term-cued recall of the definitions. In Experiment 1, some participants made judgments after reading the texts, whereas others overtly attempted retrieval of each definition before making judgments. In Experiment 2, all participants had pre-judgment recall, and some also scored the correctness of the pre-judgment responses. Accuracy was greater for term-specific than global judgments, but only when pre-judgment recall was required. Term-specific accuracy was also constrained by accessing incorrect information. We argue that TAM is not a viable explanation of accuracy and discuss how to improve judgment accuracy.

© 2005 Elsevier Inc. All rights reserved.

**Keywords:** Metacomprehension; Judgment accuracy; Metamemory; Transfer-appropriate-monitoring; Accessibility

## Introduction

Accurately monitoring how well newly studied material has been learned is critical for the effective regulation of learning (for a review, see Dunlosky, Hertzog, Kennedy, & Thiede, in press). For instance, consider two students who are studying definitions of key terms that appear within paragraphs of a classroom textbook. After studying each paragraph, they make a prediction that

reflects their confidence in correctly recalling the content of the paragraph, which is referred to as a *global* judgment because it pertains to memory for the entire paragraph. The students are then tested on their memory for the definitions, and each student's judgment accuracy is operationalized as the correlation between his or her own judgments and criterion memory performance across paragraphs. In this case, the student with the best judgment accuracy can better regulate learning, presumably because he or she will be able to focus restudy on the less well learned paragraphs (Thiede, Anderson, & Theriault, 2003). Unfortunately, the accuracy of people's global judgments reported in previous research has usually been

\* Corresponding author.

E-mail address: [jdunlosk@kent.edu](mailto:jdunlosk@kent.edu) (J. Dunlosky).

quite low, with correlations rarely exceeding +.40 (for a review, see Maki, 1998a, and for exceptions, see Thiede et al., 2003; Weaver & Bryant, 1995). What factors constrain the accuracy of metacomprehension judgments, and how can judgment accuracy be improved?

#### *Hypotheses for metacomprehension accuracy*

The present research is motivated by two hypotheses for explaining the low levels of metacomprehension accuracy. The two hypotheses—*transfer appropriate monitoring* and *accessibility*—have been empirically tested in previous research focused on explaining the accuracy of people's judgments of learning for simple materials, such as paired associates, letter strings, and individual words (e.g., Dunlosky & Nelson, 1997; Koriat, 1997; Weaver & Kelemen, 2003). By contrast, with few exceptions, these hypotheses have not been applied to understanding people's judgments of text learning. Accordingly, in the next two sections, we describe each of the hypotheses in turn and highlight how each accounts for poor metacomprehension accuracy.

#### *Transfer appropriate monitoring*

Transfer appropriate monitoring has been a particularly popular hypothesis because of its intuitive appeal as well as its origins in prominent hypotheses of memory, such as transfer appropriate processing (e.g., Roediger III, Weldon, & Challis, 1989) and material-appropriate processing (e.g., McDaniel, Einstein, Dunay, & Cobb, 1986). According to the transfer-appropriate-monitoring (TAM) hypothesis, the accuracy of people's judgments of memory are a direct function of the match between the properties of the judgment context and properties of the test context (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Dunlosky & Nelson, 1997; Glenberg, Sanocki, Epstein, & Morris, 1987; Kennedy & Nawrocki, 2003; Maki & Serra, 1992; Weaver & Kelemen, 2003). The relevant properties differ somewhat across specific variants of the TAM hypothesis. For instance, Dunlosky and Nelson (1997) argued for the importance of matching the nominal context of the judgment and test, whereas Begg et al. (1989) argued for the importance of matching the processes of the judgment and test. Neither variant of TAM has fared well in accounting for data relevant to the accuracy of metamemory judgments for paired associates (Dunlosky & Nelson, 1997; Weaver & Kelemen, 2003). However, TAM may be important for understanding metacomprehension accuracy. Relevant to this point, Weaver and Kelemen (2003) stressed that "it is possible that support for TAM may yet emerge using more complex stimulus materials such as passages of text. Text materials permit a wider range of encoding strategies and processing during judgment and test; this might increase the importance of matching processing at these times" (p. 106).

TAM may account for poor metacomprehension accuracy because of the apparent process mismatch that arises from the standard procedure for obtaining judgments of text learning. More specifically, a participant first reads a lengthy text (e.g., between 200 and 400 words) and then is asked "How well will you be able to answer questions about this text?" This prompt elicits a variety of processes. For instance, participants may evaluate their familiarity with the topic domain of the text, or they may attempt to quickly recall a brief portion of the text in the moments before making the judgment (Maki, 1998a; Morris, 1990). Such processes, however, are not likely to closely match the processes elicited by the prompts of the test questions.

According to TAM, if the match is increased between the processes that are elicited by the judgment and test, accuracy will increase. To evaluate this prediction, Glenberg et al. (1987) used a different kind of prompt for the judgments. After reading a text, participants first answered a practice test question about the text and then made a metacomprehension judgment. After the texts had been judged, criterion tests were administered. Unbeknownst to participants, the practice test questions preceding the judgments were either identical to the criterion tests, a paraphrase of the criterion tests, or unrelated to the criterion tests. When the tests were identical, the processes elicited prior to making the judgments and criterion tests would match quite well and would certainly generate a closer process match as compared to the conditions involving related or unrelated tests. Thus, metacomprehension accuracy should be greater for the identical condition than for the other two conditions. Across three experiments, accuracy was typically greater when the criterion tests were identical to the pre-judgment tests than when they were related or unrelated. Nevertheless, even when the pre-judgment tests were identical to the criterion tests, accuracy never exceeded +.40.

#### *Accessibility hypothesis*

Although TAM may provide a partial account of poor metacomprehension accuracy, the low level of accuracy demonstrated by Glenberg et al. (1987) when the process match was identical suggests that other factors may also constrain accuracy. Koriat (1993, 1995) first proposed the *accessibility hypothesis* to account for the accuracy of feeling-of-knowing judgments (for a tutorial, see Metcalfe, 2000). According to this hypothesis, people's judgments are inferential in nature, and in particular, such inferences are based on the total amount of information accessed immediately prior to making each judgment. An important assumption is that people often do not (or cannot) evaluate whether the information accessed is correct or incorrect, so only the quantity and not the quality of the accessed information is

expected to influence the metacognitive judgments.<sup>1</sup> Judgment accuracy then is a function of the empirical relationship between the amount accessed and criterion performance. If much of the accessed information is incorrect and hence not predictive of criterion performance, then the accuracy of the judgments will suffer. Thus, in contrast to TAM, the processes that precede judgments do not directly drive accuracy, but instead it is the output from those processes (and in particular, the output from retrieval processes) that is responsible for judgment accuracy.

The accessibility hypothesis provides a straightforward explanation for low metacomprehension accuracy. Namely, participants make global judgments quite quickly, and when doing so, they at best attempt to access only a portion of the to-be-judged text. Because this momentarily accessed information would not be highly diagnostic of criterion performance, accuracy tends to be low.

#### *Current approach to evaluating the two hypotheses*

TAM and the accessibility hypothesis have important differences. For TAM, the two observables that comprise judgment accuracy (judgments and criterion test performance) are merely by-products of the processes that are triggered by the judgment and test prompts, respectively. Thus, as the match between the processes increase, the judgments and criterion test performance become increasingly aligned. For the accessibility hypothesis, the judgments are not direct by-products of underlying processes. Instead, how people use the output from those processes to construct a metacognitive judgment has a well-defined influence on accuracy. Because people's judgments are inferential in nature, they can be misled when incorrect information is accessed. Thus, even if the processes that precede the judgment and test are identical, predictive accuracy can still be quite low if the output from processes preceding one's judgments is often incorrect. Of course, the two hypotheses are not mutually exclusive, and they make some shared predictions. Next, we describe one of these shared predictions and a recent test of it, and in doing so, we introduce the methods that were used in the current research to further evaluate both hypotheses.

Both hypotheses identify global judgments as centrally responsible for poor metacomprehension accu-

acy. Global judgments target a large grain size of material, with one judgment often being made for a text of over 200 words. By contrast, the grain size of a criterion test question is often small—pertaining to a single clause, concept, or definition embedded in a text. Thus, accuracy is constrained because the process match between judgment and test would be minimal (TAM) or because the output from those processes likely would have poor diagnostic validity (accessibility hypothesis). Both hypotheses predict that if the prompt for the judgment better reflects the grain size of the eventual criterion test, accuracy will improve. To evaluate this prediction, Dunlosky, Rawson, and McDonald (2002) compared the accuracy of global judgments to the accuracy of term-specific judgments, which had a grain size identical to the eventual criterion test. More specifically, college students read six paragraphs about various topics (e.g., measurement), which were each about 275 words in length. Each paragraph included four key terms (e.g., interval scale, ratio scale) along with their definitions. Student's were instructed to learn the definitions for an upcoming test. After studying a given text, they made a global judgment and then made one term-specific judgment for each of the four terms. For these judgments, a term was presented and the student was asked to predict the likelihood of recalling the correct definition for that term on an upcoming test. Immediately after these judgments were made, the criterion test was administered. For these tests, each term was presented individually, and participants were asked to type the corresponding definition.

Using each term as a prompt for the judgment affords a self-test in which students could attempt to retrieve the target information (i.e., a given definition). Thus, the process match was expected to be identical between judgment and test, and the output of such processes (i.e., access to definitions) was expected to be highly valid. Both hypotheses predict that accuracy will be greater for term-specific judgments than for the global judgments, and both hypotheses led us to expect term-specific accuracy would be quite high. In contrast to such expectations, however, accuracy for the term-specific judgments ( $M = .50$ ) did not reliably exceed the accuracy of the global judgments ( $M = .40$ ). Why did the accuracy of term-specific judgments not substantially exceed the accuracy of global judgments? A major goal of the present research was to answer this question in a way that allowed us to further evaluate the TAM and accessibility hypotheses.

One possible reason why term-specific judgments failed to significantly enhance accuracy concerns the assumption that using term-specific prompts would elicit a covert retrieval attempt of the target definition. However, the student participants may often have been cognitively lazy and instead based the judgment on a nonanalytic, gut feeling about the memory for the definition, which itself may be informed by familiarity with the term-specific prompt (Reider, 1987). The idea here is that

<sup>1</sup> Access refers more generally to not only the amount of information accessed prior to making a judgment but also to the speed of access, with quicker access speeds leading to greater degrees of confidence in one's memory. Our analyses in the present research relies on the amount accessed and not on access speed. For research relevant to the latter measure, see Benjamin, Bjork, and Schwartz (1998), Matvey, Dunlosky, and Guttentag (2001), and Morris (1990).

for at least some students, presenting the key term as the judgment prompt does not elicit a covert retrieval attempt for its definition. In fact, the idea that such prompts do not ensure complete retrieval attempts has recently received support (Kelemen, 2000; Son & Metcalfe, in press). Importantly, if the term-specific prompts did not consistently elicit retrieval attempts of the target definitions, the outcomes from Dunlosky et al. (2002) would not have provided a fair test of the two hypotheses. Accordingly, in both experiments presented in the present article, we re-evaluated the hypotheses by having some participants overtly attempt to recall each definition prior to making the term-specific judgments. By employing this method, we ensured that the process match was identical because both the prompt for the judgment and the test now demanded a retrieval attempt of the sought-after definition. If the prompts for term-specific judgments did not consistently elicit such retrieval attempts in our original research, then both hypotheses predict that term-specific accuracy will be greater for individuals who must attempt pre-judgment recall than (a) for the accuracy of global judgments and (b) for the accuracy of term-specific judgments for individuals who are not required to attempt pre-judgment recall.

As important, this pre-judgment recall and monitoring (PRAM) method (Nelson, Narens, & Dunlosky, 2004) provides other evidence for explaining the source of judgment inaccuracy that has direct implications for empirically evaluating the two hypotheses. We describe two important cases here. Consider the relationship between pre-judgment recall performance and criterion test performance. These two observables are the most relevant measures of the underlying processes, and hence their correlation reflects the degree of match between the processes that precede the judgment and the criterion test. According to TAM, judgment accuracy is expected to match the correlation between pre-judgment recall and criterion recall. In contrast to TAM, the accessibility hypothesis indicates that judgment accuracy may be significantly lower than the correlation between pre-judgment recall and criterion recall, because people's judgments do not reflect the quality of what is accessed during pre-judgment recall. For instance, if participants sometimes recall incorrect definitions prior to making the term-specific judgments, the accessibility hypothesis makes the counterintuitive prediction that participants will still judge that they will recall the correct definition. If so, according to the accessibility hypothesis, judgment accuracy is expected to match the correlation between pre-judgment recall and the term-specific judgments.

### *Overview of experiments*

In two experiments, we further evaluated TAM and the accessibility hypothesis, and in so doing identified

factors that constrain the accuracy of metacomprehension judgments. In Experiment 1, all participants made global judgments and term-specific judgments. Some participants performed pre-judgment recall whereas others did not. To foreshadow, although both the TAM and accessibility hypotheses received some empirical support, the most evident factor undermining the accuracy of term-specific judgments was the misleading influence of accessing incorrect information. Thus, in Experiment 2, we explored whether participants were able to sidestep this misleading influence by directly evaluating the quality of information accessed prior to making term-specific judgments.

## **Experiment 1**

### *Method*

#### *Design, participants, and materials*

One hundred and three students from UNCG participated to partially fulfill a requirement for an introductory psychology course. Two variables were manipulated: whether participants were prompted to retrieve the definitions prior to making the term-specific judgments (pre-judgment recall vs. no pre-judgment recall) and the study-prediction lag (minimal for immediate judgments and relatively long for delayed judgments).<sup>2</sup> Both variables were manipulated between participants, who were randomly assigned to groups by order of appearance. Number of participants included in the delayed groups were 26 (pre-judgment recall) and 27 (no pre-judgment recall), and the number of participants in the immediate groups were 24 (pre-judgment recall) and 26 (no pre-judgment recall).

Seven expository texts (one sample and six critical) were taken from introductory-level textbooks from various undergraduate courses (from Dunlosky et al., 2002). Texts were between 271 and 281 words long, with Flesch-Kincaid scores ranging from grade levels 10 to 12. Each text contained four key terms (presented in capital letters), and each term was immediately followed by a one-sentence definition. A sample text is presented in the Appendix A. Macintosh computers presented all materials and recorded responses.

<sup>2</sup> Concerning the study-prediction lag, only immediate term-specific judgments were used in our original work (Dunlosky et al., 2002), so delayed term-specific judgments were included to explore whether they would support higher levels of predictive accuracy (cf. the delayed-JOL effect, Nelson & Dunlosky, 1991). To foreshadow, because the study-prediction lag had no influence on our central measures, we discuss this factor minimally throughout the remainder of this article.

### Procedure

Participants were instructed to read several texts, to make predictions about subsequent test performance, and then to complete a cued-recall test for the keyword definitions in the texts. Before beginning the critical study trial, participants practiced each task with a sample text and test questions to familiarize them with the kind of text and recall tests they would be presented with before they began the critical trials.

For each participant, the critical texts were presented in random order, and each text was presented individually for self-paced study. Participants terminated study of a given text with a key press. For those who made immediate judgments, immediately after reading a given text they were presented the prompt for a global judgment: “How well will you be able to complete a test over this material? 0 = definitely would not be able, 20 = 20% sure I will be able, 40 = 40% sure... 100 = definitely will be able.” After making this judgment, they made four term-specific predictions. For each key term in the text, they were asked, “How well do you think you will be able to define {key term}?” These prompts were presented one at a time in random order, along with the rating scale described above. For participants who made delayed judgments, they first studied all of the texts, and then made predictions for each text. In particular, participants made a global judgment for a given text, which was prompted by the text title and the rating scale described above. They then made each of the term-specific judgments for the same text.

For groups who had pre-judgment recall, immediately after making the global judgment for a text, participants were asked “to practice recalling the definitions from the text you just read” and were prompted with the word “Define” along with a key term (e.g., “Define: Ratio Scale”) along with a field in which they were to type in the definition of that term. They were instructed to type the definition, and no time limit was imposed.

Participants first attempted to define each of the four key terms for a given text, and then they made a term-specific judgment for each one.

After the term-specific judgments (either immediate or delayed) were collected for a given text, participants had a cued-recall test for each of the four key terms. We chose this minimal retention interval because it has produced the greatest level of predictive accuracy for text in previous research (Maki, 1998b) and hence allowed us to estimate the maximal levels of accuracy supported by term-specific judgments. More important, the minimal retention interval would support the closest process match and hence provide relatively optimal conditions for evaluating the TAM hypothesis. For each definition, the key term appeared individually along with a field in which participants were to type in the definition of that term.

### Results

All differences declared as reliable had  $p < .05$ .

#### Judgment magnitudes and cued-recall performance

Mean values of performance measures have less significance for evaluating TAM and the accessibility hypothesis. Nevertheless, we begin by reporting cued-recall performance and judgment magnitudes because they are the constituent components of judgment accuracy and other derived measures that are pertinent to evaluating the hypotheses.

*Magnitude of the judgments.* For each individual, we computed the mean global judgment and mean term-specific judgment across texts (Table 1). A 2 (study-prediction lag: immediate vs. delayed)  $\times$  2 (pre-judgment recall: presence vs. absence)  $\times$  2 (kind of judgment: global vs. term-specific) ANOVA revealed only a main effect of lag,  $F(1,99) = 11.4$ ,  $MSE = 663.5$ . Mean judgments

Table 1  
Magnitude of global and term-specific judgments and recall performance on the pre-judgment test and criterion test

Group	Kind of judgment		Recall test	
	Global	Term-specific	Pre-judgment	Criterion
<i>Experiment 1</i>				
Immediate judgments				
Pre-judgment recall	41.5 (3.2)	42.4 (3.2)	43.6 (3.9)	45.4 (4.0)
No pre-judgment recall	51.5 (4.1)	52.8 (4.2)	—	51.6 (5.3)
Delayed judgments				
Pre-judgment recall	34.7 (3.5)	36.5 (3.8)	30.0 (3.6)	25.2 (3.7)
No pre-judgment recall	34.8 (3.5)	34.2 (3.7)	—	17.8 (3.4)
<i>Experiment 2</i>				
Self-score	45.7 (3.6)	48.5 (3.8)	30.1 (3.3)	29.3 (3.2)
No self-score	39.3 (3.3)	39.7 (3.7)	29.6 (3.4)	30.4 (3.4)

Note. Entries within parentheses are corresponding standard errors of each mean. In Experiment 2, both groups attempted pre-judgment recall, and the corresponding mean magnitude of the self-score rating was 53.5 ( $SEM = 3.7$ ).

decreased with greater study-prediction lag. No other main effects or interactions were reliable,  $F_s < 2.5$ ,  $MSEs < 664.0$ .

**Cued recall performance.** Each cued recall response was scored using a gist criterion. Specifically, each definition consisted of 2–4 idea units, and credit was given for responses that expressed these ideas regardless of whether the response was verbatim or a paraphrase of the original text. Partial credit was given if a response expressed some but not all of the idea units contained in a definition. Mean recall across participants for the pre-judgment tests and the criterion tests are presented in Table 1.

First, the mean percentage of *pre-judgment* recall was reliably greater for participants who attempted recall immediately after reading than for those who attempted recall at a delay,  $t(48) = 3.30$ . Second, to evaluate the effects of the two variables on *criterion* recall performance, we conducted a 2 (pre-judgment recall)  $\times$  2 (study-prediction lag) analysis of variance (ANOVA). A reliable main effect occurred for the study-prediction lag,  $F(1,99) = 41.9$ , with criterion recall being greater immediately after study than after a delay. The main effect for pre-judgment recall,  $F(1,99) = 0.02$ , and the interaction,  $F(1,99) = 2.69$ ,  $MSEs = 0.05$ , were not reliable.

#### Predictive accuracy

Two kinds of accuracy have been investigated in the literature, relative accuracy and absolute accuracy. As discussed in the Introduction, relative accuracy is measured by correlating an individual's judgments with his or her criterion test performance. Absolute accuracy is evaluated by comparing the magnitude of the judgments with the magnitude of criterion test performance. Because relative accuracy is the focus of the present

research, we highlight these results in the text. Nevertheless, term-specific judgments have not been widely investigated, so we also report absolute accuracy in Appendix B for archival purposes.

To estimate relative accuracy, we computed a Goodman–Kruskal gamma correlation between each individual's judgments and criterion recall performance (Nelson, 1984). For global judgments, the correlations involved six dyads of observations (i.e., the global judgment and recall performance for each of the six texts); for term-specific judgments, the correlations involved 24 dyads of observations (i.e., the term-specific judgments and recall performance for the 24 critical definitions). Means across individual correlations are presented in the two left-most columns of Table 2.

A 2 (study-prediction lag)  $\times$  2 (pre-judgment recall)  $\times$  2 (kind of judgment) ANOVA revealed a main effect for pre-judgment recall,  $F(1,93) = 4.37$ ,  $MSE = 0.24$ . The main effect of lag and the pre-judgment recall  $\times$  study-prediction lag interaction were not reliable [ $F(1,93) = 2.52$ ,  $MSE = 0.24$ , and  $F(1,93) = 2.89$ ,  $MSE = 0.24$ , respectively]. The main effect of kind of judgment was reliable,  $F(1,93) = 27.7$ ,  $MSE = 0.11$ , with accuracy being greater for term-specific than global judgments. The two-way interactions involving kind of judgment were not reliable,  $F_s < 2.20$ ,  $MSEs = 0.11$ . Finally, a reliable three-way interaction qualified the main effects,  $F(1,93) = 12.3$ ,  $MSE = 0.11$ . To follow-up this interaction, we conducted two separate series of analyses to evaluate the two predictions of the hypotheses discussed in the Introduction. For the first analysis, we compared the accuracy of term-specific and global judgments to evaluate the prediction that when participants must attempt pre-judgment recall, accuracy will be greater for term-specific than global judgments. In the next analyses, we evaluated whether term-specific

Table 2

Correlations between judgments and criterion recall (judgment accuracy) and among pre-judgment recall, criterion recall and term-specific judgments

Group	Judgment accuracy		Pre-J Rec and Criterion Rec	Pre-J Rec and Term-specific
	Global	Term-specific		
<i>Experiment 1</i>				
Immediate judgments				
Pre-J recall	.41 (.09)	.73 (.04)	.88 (.02)	.71 (.03)
No pre-J recall	.52 (.12)	.57 (.05)	—	—
Delayed judgments				
Pre-J recall	.52 (.09)	.64 (.05)	.94 (.02)	.63 (.06)
No pre-J recall	.05 (.13)	.57 (.08)	—	—
<i>Experiment 2</i>				
Self-score	.34 (.09)	.57 (.05)	.89 (.05)	.57 (.05)
No self-score	.23 (.09)	.61 (.06)	.89 (.05)	.62 (.06)

*Note.* Judgment accuracy is the mean across individual participant correlations between judgments and criterion recall. Pre-J Rec, pre-judgment recall. Criterion Rec, criterion recall. Term-specific, term-specific judgment. Entries within parentheses are corresponding standard errors of each mean. In Experiment 2, both groups attempted pre-judgment recall.

accuracy was positively influenced by pre-judgment recall attempts.

To establish a replication of our earlier research, we began by comparing the accuracy of global judgments and term-specific judgments when participants were not forced to make pre-judgment recall attempts. As in Dunlosky et al. (2002), accuracy was not reliably different for global and term-specific judgments when they were made immediately after study,  $t(25) = 0.40$ . For delayed judgments, accuracy was reliably greater for term-specific than global judgments,  $t(21) = 4.98$ ,  $p < .01$ . Note, however, that this latter effect is primarily due to the low accuracy of delayed global judgments, which replicates Maki (1998b). Most important, forcing participants to recall definitions boosted the accuracy of term-specific judgments above the accuracy of global judgments. In particular, accuracy was reliably greater for term-specific judgments than for global judgments when they were immediate with pre-judgment recall,  $t(22) = 3.87$ ,  $p < .01$ , and the same trend was evident for delayed judgments, although it was not statistically reliable,  $t(25) = 1.50$ .

To examine whether pre-judgment recall influenced the accuracy of term-specific judgments, we conducted a 2 (study-prediction lag)  $\times$  2 (pre-judgment recall) ANOVA, which did not reveal a reliable main effect for study-prediction lag or a reliable interaction,  $F_s < 1.0$ ,  $MSEs = 0.07$ . The effect of pre-judgment recall was reliable,  $F(1,94) = 4.79$ ,  $MSE = 0.07$ , indicating that pre-judgment recall boosted the accuracy of term-specific judgments.

In summary, when participants were forced to make a retrieval attempt of each target definition, term-specific accuracy reliably improved regardless of the study-prediction lag and was reliably boosted above global accuracy for immediate judgments. These outcomes are consistent with the idea that even though a term-specific judgment affords an opportunity to assess memory vis-à-vis a retrieval attempt, this affordance does not ensure that such an attempt will be made.

#### *Relations among pre-judgment recall, term-specific judgments, and criterion recall*

*Correlational analyses.* To explore more detailed predictions of the TAM and accessibility hypotheses, we report intra-individual correlations between pre-judgment recall and criterion test performance and between pre-judgment recall and the term-specific judgments. Recall that TAM predicts that the accuracy of term-specific judgments will be restricted by the former correlation, which reflects the match between processes that support term-specific judgments and criterion recall. The accessibility hypothesis predicts that the accuracy of term-specific judgments will be restricted by the latter correlation, because participants may judge a definition to be known even if an incorrect definition is accessed during pre-judgment recall.

In Table 2, we report the correlations between our three focal variables: pre-judgment recall, term-specific judgments, and criterion recall. Several outcomes are noteworthy. First, the correlation between pre-judgment recall and criterion recall was quite high regardless of the study-prediction lag, which is consistent with the assumption that the processes that precede the judgments and the criterion test were closely matched. In contrast to TAM, however, the accuracy of term-specific judgments was reliably lower than this correlation both for immediate judgments,  $t(23) = 3.46$ , and for delayed judgments,  $t(25) = 6.17$ . Second, the correlations between pre-judgment recall and the term-specific judgments were nearly identical to the level of accuracy for immediate and delayed judgments,  $t_s < 0.60$ . These outcomes provide more competitive support for the accessibility hypothesis than for TAM.

*Content analyses of pre-judgment cued recall.* People's term-specific judgments did not fully capitalize on the high diagnosticity of output from pre-judgment recall, as suggested by the less-than-perfect correlations between pre-judgment recall and the term-specific predictions. Why? According to the accessibility hypothesis, inappropriate use of a cue could occur if students occasionally retrieved incorrect information about a term during pre-judgment recall. In this case, although incorrect recall (i.e., a commission error) would be diagnostic of incorrect criterion recall, participants' judgments would be inappropriately high because some information—albeit incorrect—had been accessed prior to making the judgment.

To evaluate this possibility, we plotted the magnitude of term-specific judgments and criterion recall performance as a function of the content of participants' pre-judgment recall. The content of pre-judgment recall included four kinds of response: *omissions*, in which nothing was output during a pre-judgment recall attempt; *commissions*, in which the output was objectively incorrect; *partially correct*, in which the output contained some but not all of the correct definition; and *correct* answers, in which the output entirely specified the gist of the correct definition. For each of these kinds of response, we computed means across individual term-specific judgments and criterion recall, which are plotted in Fig. 1.

Most important and consistent with expectations from the accessibility hypothesis, when participants made commission errors during pre-judgment recall, the term-specific judgments overestimated criterion recall for both immediate judgments,  $t(21) = 3.14$ , and delayed judgments,  $t(24) = 6.01$ . For correct recall, the reverse occurred, with judgments underestimating criterion recall, for both immediate judgments,  $t(23) = 6.15$ , and delayed judgments,  $t(23) = 3.56$ .

The magnitudes of term-specific judgments varied as a function of the content of pre-judgment recall,

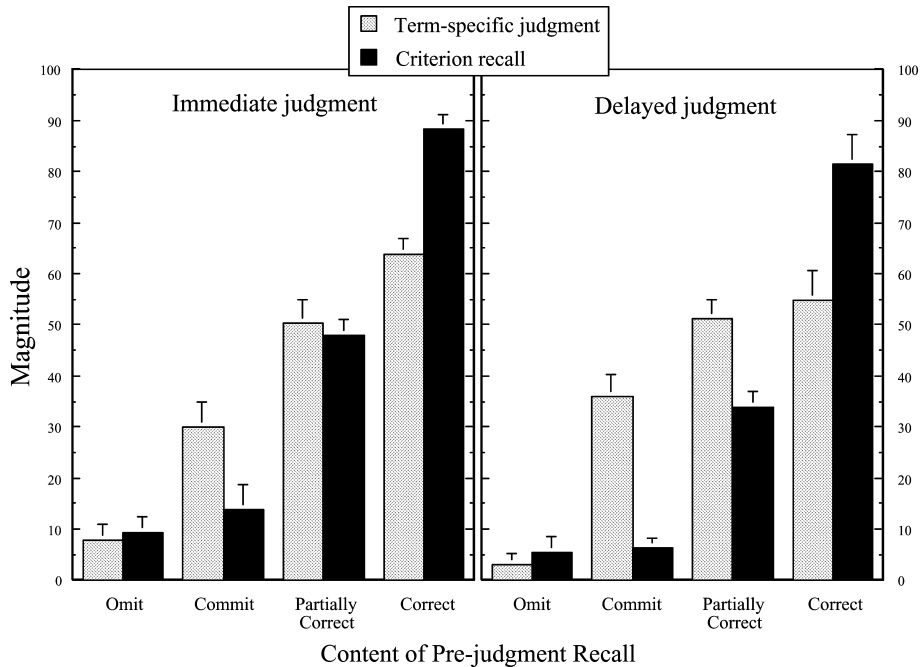


Fig. 1. Mean term-specific judgments and mean criterion recall performance conditionalized on the content of pre-judgment recall. Omit, omission during pre-judgment recall; Commit, response during pre-judgment recall was entirely incorrect; Partially correct, response included some but not all of the correct definition; Correct, response included all of the correct definition. See text for details on scoring. Error bars are the corresponding standard error of the mean.

$F(3,114) = 93.8$ ,  $MSE = 238.6$ . Most obvious, term-specific judgments were lowest for omission errors when no information had been output, which is consistent with the accessibility hypothesis. However, a pattern also emerged that was not anticipated by the accessibility hypothesis. Namely, term-specific judgments following a pre-judgment recall attempt were related to the qualitative content of the pre-judgment recall. For both immediate and delayed judgments, follow-up  $t$  tests (with criterion  $\alpha = .008$  as adjusted by the Bonferroni correction) indicated the following order of term-specific judgments: commissions < partially correct < correct (although the latter comparison between partially correct and correct was not reliably significant for delayed judgments,  $p = .38$ ). This pattern suggests that individuals' judgments may be based not only on amount accessed but also on the quality of information that is accessed.

#### *Relations among the quality and quantity of pre-judgment recall and term-specific judgments*

This particular conclusion is tentative, however, because the quantity and the quality of the output of pre-judgment recall were partially confounded—that is, less information was accessed when participants responded with commission errors than correct responses. Could the lower term-specific judgments after commission errors be accounted for by sheer differences

in the amount accessed, or do term-specific judgments capitalize on quality above and beyond the amount accessed? Put differently, do higher term-specific judgments arise for correct responses than for commissions because people in part evaluate the quality of accessed information, or does it arise solely from differences in the amount of information accessed?

To empirically evaluate the latter possibility, we correlated the number of words recalled during pre-judgment recall with the content status of that recall (i.e., commission error vs. correct response). Commissions were coded as 0, and correct responses were coded as 1, so that the correlation would be positive if the amount accessed was greater for correct responses than for commissions. Means across intra-individual correlations are presented in the left-most columns of Table 3. (We present both gamma and Pearson  $r$  correlation coefficients for reasons discussed below.) The correlations tended to be positive (albeit not always reliably so), which empirically establishes that more information was accessed when people responded with correct answers than with commission errors.

This difference in accessibility could account for the relationship between content status of pre-judgment recall (commissions vs. correct responses) and the magnitude of term-specific judgments. This possibility was evaluated in two steps. First, we correlated the content

Table 3  
Relations between content of pre-judgment recall, quantity of pre-judgment recall, and term-specific judgments

Group	Correlation between content and				Correlation between content and TS judgment	
	Total # words		TS Judgment		Partialled on total # words	
	Gamma	<i>r</i>	Gamma	<i>r</i>	Gamma	<i>r</i>
<i>Experiment 1</i>						
Immediate judgments	.33 (.12)	.26 (.08)	.84 (.05)	.60 (.05)	.83 (.06)	.56 (.05)
Delayed judgments	.21 (.12)	.09 (.08)	.46 (.13)	.29 (.08)	.51 (.14)	.27 (.08)
<i>Experiment 2</i>						
Self-score	.30 (.08)	.20 (.06)	.57 (.08)	.37 (.06)	.58 (.09)	.35 (.07)
No self-score	.38 (.07)	.24 (.05)	.72 (.08)	.43 (.06)	.76 (.08)	.41 (.06)

*Note.* Content, content of pre-judgment recall included only commission errors (scored as 0) or correct responses (scored as 1), so positive correlations signify that correct responses were longer (total # words) or received higher term-specific judgments. Total # words, total number of words in the response to pre-judgment recall. TS judgment, term-specific judgment. All correlations were computed within each participant, and cell entries are means across correlations (with standard errors of the mean in parentheses).

status of pre-judgment recall (commissions vs. correct responses) and term-specific judgments (which was expected to be positive in correspondence with the qualitative effects presented in Fig. 1). Second, we compared this correlation to the same correlation after partialling out variance associated with the total number of words recalled. If amount accessed is solely responsible for the effect of content on the judgments, partialling on the total number words recalled will yield a correlation of 0 between content and the term-specific judgments.

The correlation between content (commission = 0, correct response = 1) and the term-specific judgments are reported in the two middle columns of Table 3. These correlations were positive, which reflected the qualitative effect of the content status (commissions vs. correct) of responses on the judgments. The gamma correlations were partialled on total words recalled by first separating responses into two bins based on the amount recalled. Two bins (instead of more than two) were used to increase the number of comparisons that would be used to estimate the partial gamma. Because substantial variability of the total number of words may still remain within each bin, we also estimated the corresponding values using a Pearson *r* correlation. As is evident from inspection of Table 3, both gamma and *r* led to the same conclusions: the relationship between judgments and the content status of the responses remained the same after the total number of words was partialled out. Thus, when making term-specific judgments, people apparently base their judgments on more than just the amount of information accessed.

### Discussion

Results from Experiment 1 suggest that the prompt for term-specific judgments does not always lead to a complete retrieval attempt for a target definition, as indicated by the improvement in accuracy of term-specific judgments when participants were forced to attempt pre-

judgment recall. This outcome is consistent with both TAM and the accessibility hypothesis. Nevertheless, the overall pattern of findings provided more competitive support for the accessibility hypothesis than for TAM. Most relevant is that when making term-specific judgments, people were apparently misled when they had accessed incorrect definitions. Whereas this effect cannot be accounted for by the passive mechanism of TAM, the influence of incorrect access on judgments is consistent with the accessibility hypothesis. Finally, further analyses of the quality and quantity of output (Table 3) suggest that the judgments do not solely result from accessibility but may also be influenced by an analyses of the quality of what had been accessed.

### Experiment 2

To revisit, the accessibility hypothesis assumes that people evaluate the quantity of accessed information rather than its quality. However, Experiment 1 provided tentative evidence that individuals may also assess the quality of their output when making judgments. Thus, a major goal of Experiment 2 was to further explore the extent to which individuals can evaluate the quality of accessed information. Although analyses reported in Table 3 suggest that participants had some skill at evaluating quality, the prediction-based method used in Experiment 1 may underestimate how well individuals actually can evaluate quality. That is, asking participants to *predict* future performance may have reduced the likelihood that they would engage in an analysis of quality, even though they are even more capable of doing so. For instance, results from other research have demonstrated that when people are asked to predict future performance, the judgments are influenced by factors such as an individual's theory of retention (Rawson, Dunlosky, & McDonald, 2002), the structure of the anticipated test (Thiede, 1996), and so forth. By basing the predictive

judgments on factors that are extrinsic to the target information, individuals may under-weight an analysis of the quality of accessed information when constructing their predictive judgments.

To address this issue, we examined whether a direct, analytic evaluation of the content of information accessed prior to making term-specific judgments would improve judgment accuracy. In particular, we had two groups of participants make term-specific judgments, which were delayed and were preceded by pre-judgment recall attempts. Before making their term-specific judgments, one group first explicitly rated the quality of the information retrieved during the pre-judgment recall attempt. That is, we had them score their own recall output. If participants are capable of accurately evaluating accessed information yet fail to fully use such evaluations when predicting future performance, we expected that (a) participants will self-score commission errors as incorrect, and hence (b) the accuracy of term-specific judgments will be greater for those who self-scored their pre-judgment recall than for those who did not.

### Method

#### *Design, participants, and materials*

Seventy-nine students from UNCG participated to partially fulfill a requirement for an introductory psychology course. One between-subjects variable was manipulated: whether or not participants scored their pre-judgment recall prior to making term-specific judgments. Forty participants were assigned to the self-score group and 39 were assigned to the group who did not self-score. The materials and computers were identical to those used in Experiment 1.

#### *Procedure*

The procedure was identical to the one used in Experiment 1, with the following exceptions. First, all participants made delayed judgments and attempted pre-judgment recall. Second, one group of participants also scored their own pre-judgment recall. Specifically, immediately after typing their response to a particular term-specific prompt for pre-judgment recall, participants answered the following question: "If the correctness of the definition you just wrote was being graded, do you think you would receive: no credit, partial credit, or full credit?" The unambiguous three-point scale thus gave participants the opportunity to demonstrate their ability to identify commission errors by choosing "no credit." In the analyses, the points of the rating scale were scored as 0% for responses of "no credit," 50% for "partial credit," and 100% for "full credit," which best reflected the corresponding scale used for the experimenter scoring of recall responses.

### Results

#### *Judgment magnitudes and cued-recall performance*

For each individual, we computed the mean global judgment and mean term-specific judgment across texts, which are presented in the bottom half of Table 1. A 2 (self-score vs. no self-score)  $\times$  2 (kind of judgment) ANOVA did not reveal any reliable effects: for self-scoring,  $F(1, 77) = 2.34$ ,  $MSE = 964.3$ , for kind of judgment,  $F(1, 77) = 1.64$ ,  $MSE = 60.7$ , or for the interaction,  $F(1, 77) < 1.0$ ,  $MSE = 60.7$ . The magnitude of the self-score rating was 53.5 ( $SEM = 3.7$ ).

As evident from inspection of Table 1, self-scoring pre-judgment recall did not reliably influence performance on either pre-judgment recall,  $t(77) = 0.10$ , or on criterion recall,  $t(77) = 0.23$ .

#### *Predictive accuracy*

We estimated the relative accuracy of the term-specific judgments and the global judgments as in Experiment 1. (Estimates of absolute accuracy are reported in Appendix B.) Means across individual correlations are presented in the left-most columns of Table 2. A 2 (self-score vs. no self-score)  $\times$  2 (kind of judgment: global vs. term-specific judgment) ANOVA revealed a main effect for kind of judgment,  $F(1, 71) = 18.1$ ,  $MSE = 0.19$ , indicating that accuracy was greater for term-specific than global judgments. The main effect of self-scoring and the interaction were not reliable,  $F_s < 1.0$ .

Having participants explicitly score the quality of their pre-judgment recall did not significantly boost the accuracy of term-specific judgments. One possible explanation for this outcome is that participants' self scores were actually predictive of criterion recall, but they still did not incorporate their assessments of quality into their term-specific judgments. The data, however, are inconsistent with this explanation. The correlation between self-score ratings and criterion recall was only .64 ( $SEM = .05$ ), whereas the correlation between self-score ratings and the term-specific judgments approached unity ( $M = .95$ ,  $SEM = .02$ ,  $Mdn = .99$ ).

#### *Relations among pre-judgment recall, term-specific judgments, and criterion recall*

*Correlational analyses.* To further evaluate the TAM and accessibility hypotheses, we computed correlations between pre-judgment recall and criterion test performance and between pre-judgment recall and the term-specific judgments. In Table 2, we report the correlations between our three focal variables: pre-judgment recall, term-specific judgments, and criterion recall. Consider several outcomes. First, the correlation between pre-judgment recall and criterion recall was quite high (Median correlations were +.95 for both groups), which again confirms the assumption that the match was nearly perfect between processes that preceded the judgments and the

criterion test. In contrast to TAM, however, the accuracy of term-specific judgments was reliably lower than this correlation both for those who scored their pre-judgment recall,  $t(37)=4.37$ , and for those who did not score pre-judgment recall,  $t(36)=3.53$ . Second, the correlations between pre-judgment recall and the term-specific judgments were nearly identical to the level of term-specific accuracy for both groups,  $r_s < 1.0$ . These outcomes provide more competitive support for the accessibility hypothesis than for TAM.

Finally, the correlation between pre-judgment recall and the self-score ratings was  $.71$  ( $SEM = .05$ ), which was significantly above zero but also significantly less than  $1.0$ . This correlation indicates that although participants had some skill at evaluating the quality of output from pre-judgment recall, they obviously had some difficulty in doing so. We explore the source of this difficulty in the next section.

*Content analyses of pre-judgment recall.* People's term-specific judgments did not fully capitalize on the high validity of the output of pre-judgment recall for predicting criterion recall, even when they were forced to evaluate the quality of the output. To discover the source of participants' difficulty, we plotted the magnitude of term-specific judgments, self-scoring ratings, and criterion recall performance as a function of the content of participants' pre-judgment recall. The means across indi-

vidual values for each of the pre-judgment recall responses (omissions, commissions, partially correct, and correct) are plotted in Fig. 2.

As evident from inspection of Fig. 2, the outcomes were qualitatively similar to those from Experiment 1, even for the self-score group. First, consider term-specific judgments and criterion recall. When participants made commission errors during pre-judgment recall, the magnitude of term-specific judgments was reliably greater than the magnitude of criterion recall both for those who scored their pre-judgment recall,  $t(37)=13.6$ , and for those who did not score pre-judgment recall,  $t(38)=9.03$ . When participants correctly recalled the definitions during pre-judgment recall, the reverse occurred, with term-specific judgments reliably underestimating criterion recall for the self-score group,  $t(32)=2.27$ , and for those who did not self-score,  $t(31)=7.73$ .

Self-scoring also had a minimal influence on the qualitative pattern of the term-specific judgments across the content of pre-judgment recall. In particular, for both the self-score and no self-score groups,  $t$  tests (with criterion  $\alpha = .008$  as adjusted by the Bonferroni correction) indicated the following order of term-specific judgments: omissions < commissions < partially correct < correct (although the comparison between partially correct and correct was not reliably significant for the self-score group,  $p = .04$ ).

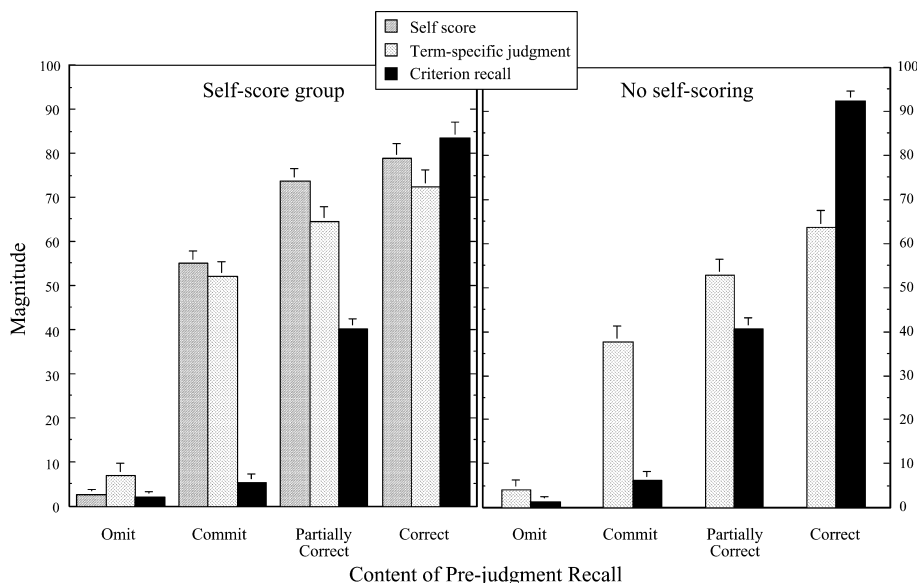


Fig. 2. Mean self-scoring rating, mean term-specific judgments, and mean criterion recall performance conditionalized on the content of pre-judgment recall. Self-scoring ratings were omitted from the right panel because those values are from a group who did not self score pre-judgment recall. Omit, omission during pre-judgment recall; Commit, response during pre-judgment recall was entirely incorrect; Partially correct, response included some but not all of the correct definition; Correct, response included all of the correct definition. See text for details on scoring. Error bars are the corresponding standard error of the mean.

### *Relations among the quality and quantity of pre-judgment recall and term-specific judgments*

Although the pattern of effects presented in Fig. 2 suggests that term-specific judgments are in part influenced by assessments of output quality, this conclusion is tentative because the quality and quantity of the output of pre-judgment recall were partially confounded. As in Experiment 1, we statistically evaluated whether the relationship between the magnitude of term-specific judgments and the content status of pre-judgment recall (i.e., commission vs. correct) could be accounted for by differences in the quantity of output. To do so, we correlated the number of words recalled during pre-judgment recall with the content status of that recall (i.e., commission error vs. correct response) and then computed this correlation partialling on the total number of words (quantity) recalled during pre-judgment recall. Means across intra-individual correlations are presented in Table 3.

Outcomes based on gamma and  $r$  led to the same conclusion. Namely, the relationship between judgments and the content status of the responses remained the same after the total number of words was partialled out. Thus, people apparently base their term-specific judgments on more than just the amount of information accessed.

### *Discussion*

Results from Experiment 2 provide a critical replication of key outcomes from Experiment 1. Pre-judgment recall boosted the accuracy of term-specific judgments over the accuracy of the global judgments, which is consistent with a prediction from both TAM and the accessibility hypothesis. Correlational analyses provided further support for the accessibility hypothesis in that participants were misled by inaccurate information—i.e., commission errors—that was accessed prior to making the term-specific judgments. Finally, even when participants were forced to evaluate the quality of their pre-judgment recall, this evaluation did not improve the accuracy of term-specific judgments.

### **General discussion**

The present research evaluated the prediction of two hypotheses that increasing the match between the prompt for metacomprehension judgments and the prompt for the criterion test will improve the predictive accuracy of the judgments. In Experiment 1, increasing the nominal match by using term-specific judgments did not produce reliably greater accuracy than that obtained by global judgments, which are standard in the field. Importantly, across both experiments, when participants were forced to recall the target definitions prior to mak-

ing term-specific judgments, the accuracy of these judgments reliably improved. In these cases, however, even though the match between the prompts for judgment and criterion test was nearly identical, the level of predictive accuracy was far from perfect, with the mean level across experiments being around  $+0.64$ . Thus, the accuracy of metacomprehension judgments can be improved by matching the prompts for the judgment and criterion test, but even an identical match does not guarantee maximal accuracy.

### *Transfer-appropriate-monitoring vs. the accessibility hypothesis*

Although TAM offers a straightforward explanation for the positive effect of pre-judgment recall on the accuracy of term-specific judgments, other evidence fails to support a critical prediction from TAM. In particular, the correlation between the outcome of pre-judgment recall and criterion recall was quite high ( $\approx 0.90$ ) across both experiments, indicating a close match between processes at judgment and criterion test. TAM predicts that predictive accuracy for the term-specific judgments will reach the same level of magnitude, yet under these circumstances accuracy was still constrained (Table 2).

Why is TAM inadequate? TAM accounts for accuracy via a *passive* mechanism that concerns matching the processes elicited by the prompt for the judgments and the criterion test. This account is inherently correlational and does not specify the underlying cause(s) of accuracy, which ultimately undermines its ability to explain predictive accuracy at all. An analogous problem with memory hypotheses that originally inspired TAM has recently been described by Nairne (2002): “The link between the encoding-retrieval match and retention, although generally positive, is effectively correlational rather than causal. Increasing the functional similarity between a cue and a target trace can improve, have no effect, or even decrease the retention performance depending on the circumstance...When we remember, we use the information at hand, in the form of retrieval cues, to make a decision about what occurred in the past. But the decision is unlikely to be based on a passive matching process” (p. 390). Likewise, when people predict future memory performance, they apparently use the information at hand. How they use this information to judge what will occur is essential. Thus, because the accuracy of judgments does not result from a passive matching process, even if the process match is identical, predictive accuracy may be perfect, nil, or even negative, depending upon how individuals use the available information to construct the judgments.

The overall pattern of data provides more competitive support for the accessibility hypothesis than for TAM. The accessibility hypothesis offers a causal explanation for predictive accuracy by describing how people

use accessed information to construct metacognitive judgments. According to the accessibility hypothesis, metacomprehension judgments are based on the amount of information accessed prior to making the judgment, such as text content that is accessed in the moments before making a global judgment (e.g., Morris, 1990) or more specific content accessed prior to making term-specific judgments. The judgments are presumed to be based on the amount of information accessed and not on the quality of the accessed information. In the present circumstances, predictive accuracy was well above chance because some of the accessed information was valid. That is, during pre-judgment recall, participants sometimes recalled nothing (omissions) or recalled the correct definitions during pre-judgment recall, which were both predictive of the eventual outcomes on the criterion test. Importantly, however, accessing incorrect information—i.e., commission errors—resulted in elevated judgments that did not reflect the eventual outcomes on the criterion test, thus constraining accuracy.

*Do term-specific judgments capitalize on quality beyond access?*

The accessibility hypothesis was proposed by Koriat (1993, 1995) to account for the accuracy of feeling-of-knowing (FOK) judgments. According to this hypothesis, the magnitude of FOK judgments are presumably based on the amount of information accessed (for alternative accounts, see Metcalfe, Schwartz, & Joaquim, 1993; Reder & Ritter, 1992). FOK accuracy then depends on the correctness of the information accessed, because people's FOK judgments are largely (if not completely) uninfluenced by the quality of information that comes to mind. An important difference exists between FOK judgments and those investigated here, which pertains to the kind of information accessed prior to making the judgments. FOK judgments typically involve predicting the likelihood of recognizing relatively small units of information, such as a recently studied letter string or a currently unrecallable answer to a trivia question. Accordingly, FOK judgments are influenced by access to partial information about these sought-after targets, which in this example may include accessing a few letters or the semantic attributes of an answer. As noted by Koriat (1995), although these "clues may not be articulate enough to support an analytic, calculated inference, they can still act en masse to produce the subjective feeling that the target is there and will be recalled or recognized in the future" (p. 312).

By contrast, information accessed prior to making term-specific judgments often consisted of a relatively well-articulated series of words, concepts, and ideas. Accessing these larger conceptual units—especially when they are elicited by a pre-judgment retrieval prompt—may support a more analytic inference based on an

assessment of the quality of the information accessed. On one hand, the magnitude of term-specific judgments were well above the floor even when the information accessed was entirely incorrect (Fig. 1), which may mean that participants often did not (or could not) engage in an analysis of the quality of the accessed information. On the other hand, evidence from both experiments also suggests that participants partly evaluated the quality of the accessed information. In particular, participants' term-specific judgments were lower for commission errors than for correctly recalled definitions. Importantly, this relationship remained even after the amount of information accessed was controlled for statistically (Table 3).

Of course, the speed of accessing information may also be influential, if access is slower for commission errors than for correct responses.<sup>1</sup> We cannot rule out this possibility with the current data, although it seems unlikely that access speed could account entirely for the sizable influence that the content of pre-judgment recall has on the judgments. At the least, the current evidence cannot rule out the possibility that students are capable of evaluating the quality of information that they access prior to making term-specific judgments. Realizing the extent of this capability, and whether it can be enhanced, awaits future research.

*Recommendations for improving metacomprehension accuracy*

The present evaluation of the two hypotheses leads to some recommendations for improving metacomprehension accuracy. First, developing prompts that match the grain size of the criterion test can improve judgment accuracy, presumably because term-specific prompts afforded retrieval attempts that led to the access of information that was relatively valid for predicting criterion performance. Matching grain size will tend to improve accuracy when it produces correct self-feedback about learning (Glenberg et al., 1987); nevertheless, even a well-chosen prompt for a judgment may produce incorrect self-feedback, which will constrain metacomprehension accuracy. Thus, matching the grain size between the prompt for judgments and the criterion test may be necessary but alone may often not be sufficient for consistently achieving high levels of accuracy. Second, because people may be misled by incorrect information that is accessed when making term-specific judgments, another avenue for improving accuracy will involve developing techniques to help individuals more accurately evaluate the quality of their knowledge. Preferably, these techniques will be easy to use and will be applicable both to a broad range of learner abilities as well as to any content domain.

Given that metacomprehension research has almost exclusively focused on global judgments, whether individuals will be able to use term-specific judgments—along with special instruction or training—to achieve

maximal accuracy is not yet known. Even so, outcomes from the present research arguably demonstrate the promise of term-specific judgments both for improving accuracy in particular and for enhancing student self-regulated learning of text materials in general.

## Appendix A. Sample text

### Gestures

Scholars who have studied body language extensively have devised a widely used system to classify the function of gestures that people use when speaking publicly. EMBLEMS are gestures that stand for words or ideas. You occasionally use them in public speaking, as when you hold up your hand to cut off applause. Emblems vary from culture to culture. The sign that stands for "a-ok" in this country refers to money in Japan, and it is an obscene gesture in some Latin American countries. ILLUSTRATORS are gestures that simply illustrate or add emphasis to your words. For example, speakers often pound on a podium to accent words or phrases. In addition, you can illustrate spatial relationships by pointing or by extending your hands to indicate width or height. ADAPTORS are a different group of gestures used to satisfy physical or psychological needs. SELF-ADAPTORS are those in which you touch yourself in order to release stress. If you fidget with your hair, scratch your face, or tap your leg during a speech, you are adapting to stress by using a self-adaptor. You use object-adaptors when you play with your keys, twirl a ring, jingle change in your pocket, or tap pencils and note cards. Finally, ALTER-ADAPTORS are gestures you use in relation to the audience to protect yourself. For instance, if you fold your arms across your chest during intense questioning, you may be subconsciously protecting yourself against the perceived psychological threat of the questioner. Whereas emblems and illustrators can be effective additions to a speech, adaptors indicate anxiety and appear as nervous mannerisms and should therefore be eliminated from public speaking habits.

## Appendix B

For archival purposes, we present levels of absolute accuracy for both kinds of judgment. For each kind of judgment, we computed a difference score between an individual's mean judgment and mean recall performance across tests. Means across individual difference scores are presented in Table B1 for both experiments.

For Experiment 1, a 2 (kind of judgment)  $\times$  2 (study-prediction lag)  $\times$  2 (pre-judgment vs. no pre-judgment recall) ANOVA revealed no reliable main effects or interactions,  $F_s < 1.0$ ,  $MSEs < 980.0$ , except for a reliable main effect of study-prediction lag,  $F(1,99) = 11.8$ ,  $MSE = 976.4$ , which indicates that absolute accuracy was greater for immediate than delayed judgments.

For Experiment 2, a 2 (kind of judgment)  $\times$  2 (self-score vs. no self-score) ANOVA did not reveal a reliable effect of kind of judgment or a reliable interaction,  $F_s < 1.7$ ,  $MSE_s < 770$ . The effect of self-scoring approached reliability,  $F(1,77) = 3.83$ ,  $MSE = 766.1$ ,  $p = .054$ .

### Absolute accuracy for global judgments and term-specific judgments

Group	Kind of judgment	
	Global	Term-specific
<i>Experiment 1</i>		
Pre-judgment recall		
Immediate judgments	-3.8 (4.4)	-3.0 (4.1)
Delayed judgments	9.5 (3.9)	11.3 (3.8)
No pre-judgment recall		
Immediate judgments	-0.1 (5.6)	1.3 (5.3)
Delayed judgments	17.0 (4.2)	16.4 (4.0)
<i>Experiment 2</i>		
Self-score	16.3 (3.3)	19.1 (3.2)
No self-score	8.9 (3.4)	9.3 (3.1)

*Note.* Absolute accuracy is the mean across individual difference scores between mean judgments and test performance. Value in parentheses is the standard error of the corresponding mean.

## References

- Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, 28, 610–632.
- Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metacognitive index. *Journal of Experimental Psychology: General*, 127, 55–68.
- Dunlosky, J., Hertzog, C., Kennedy, M. R. T., & Thiede, K. W. (in press). The self-monitoring approach for effective learning. *Cognitive Technology*.
- Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language*, 36, 34–49.
- Dunlosky, J., Rawson, A. K., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. Perfect, B. Schwartz (Eds.) *Applied Metacognition*.
- Glenberg, A., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, 116, 119–136.
- Kennedy, M. R. T., & Nawrocki, M. D. (2003). Delayed predictive accuracy of narrative recall after traumatic brain injury: Salience and explicitness. *Journal of Speech, Language, and Hearing Research*, 46, 98–112.
- 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.
- Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609–639.
- Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General*, 124, 311–333.
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of

- learning. *Journal of Experimental Psychology: General*, 126, 349–370.
- Maki, R. H. (1998a). 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: LEA.
- Maki, R. H. (1998b). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition*, 26, 959–964.
- Maki, R. H., & Serra, M. (1992). Role of practice tests on the accuracy of test predictions on text material. *Journal of Educational Psychology*, 84, 200–210.
- Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs?. *Memory & Cognition*, 29, 222–232.
- Metcalf, J. (2000). Metamemory: Theory and data. In E. Tulving & F. I. M. Craik (Eds.), *The oxford handbook of memory* (pp. 197–211). New York: Oxford University Press.
- Metcalf, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition, Memory, and Cognition. *Journal of Experimental Psychology: Learning*, 19, 851–861.
- McDaniel, M. A., Einstein, G. O., Dunay, P. K., & Cobb, R. E. (1986). Encoding difficulty and memory: Toward a unifying theory. *Journal of Memory and Language*, 5, 645–656.
- Morris, C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 223–232.
- Nairne, J. S. (2002). The myth of encoding-retrieval match. *Memory*, 10, 289–395.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95, 109–133.
- Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The delayed-JOL effect. *Psychological Science*, 2, 267–270.
- Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods*, 9, 53–69.
- Rawson, K., Dunlosky, J., & McDonald, S. (2002). Influences of metamemory on performance predictions for text. *Quarterly Journal of Experimental Psychology*, 55A, 505–524.
- Reder, L. (1987). Strategy selection in question answering. *Cognitive Psychology*, 19, 90–138.
- Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing. Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 435–451.
- Roediger III, H. L., Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 3–41). Hillsdale, NJ: Erlbaum.
- Son, L., & Metcalf, J. (in press). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition*.
- Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 49A, 901–918.
- Thiede, K. W., Anderson, M. C. M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, 95, 66–73.
- Weaver, C. A., & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition*, 23, 12–22.
- Weaver, C. A., & Kelemen, W. L. (2003). Processing similarity does not improve metamemory: Evidence against transfer-appropriate-monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1058–1065.