

Metacomprehension

A Brief History and How to Improve Its Accuracy

John Dunlosky and Amanda R. Lipko

Kent State University

ABSTRACT—*People’s judgments about how well they have learned and comprehended text materials can be important for effectively regulating learning, but only if those judgments are accurate. Over two decades of research examining judgments of text learning—or metacomprehension—has consistently demonstrated that people’s judgment accuracy is quite poor. We review recent research that has shown some success in improving judgment accuracy and then argue that the most common method used to investigate metacomprehension accuracy may inadvertently constrain it. We describe a new method that sidesteps some problems of the older method and present evidence showing how people can achieve high levels of metacomprehension accuracy.*

KEYWORDS—*metacomprehension; judgment accuracy; metacognition; text learning*

Metacomprehension refers to a person’s ability to judge his or her own learning and/or comprehension of text materials. Researchers have fervently investigated the accuracy of people’s metacomprehension judgments, because the importance of achieving high levels of metacomprehension accuracy is evident in many areas. In learning new jobs, trainees often must acquire a great deal of new information, such as about organizational structures and how to accomplish specific tasks. Mastering many hobbies, from scuba diving to guitar playing, often requires learning and understanding a great deal of new information. And in education, given national mandates to “leave no child behind,” grade-school students are expected to learn a tremendous amount of course content in a limited amount of time, and such expectations follow many students through high school and college. To meet these kinds of learning goals, people must use their time efficiently, which is where accurate metacomprehension can play a key role.

In particular, if people can judge what material they have learned well and what they have not, they can focus their attention just on unlearned information. If their metacomprehension accuracy is poor, however, they will not be able to use their judgments to appropriately guide learning (Dunlosky, Hertzog, Kennedy, & Thiede, 2005). For instance, Thiede, Anderson, and Therriault (2003) had college students study six texts on topics such as Norse settlements and naval warfare. After an initial round of self-paced study, the students were asked to judge how well they understood each text and later answered six initial questions about each one. Importantly, the researchers influenced how different groups made the metacomprehension judgments, so that one group’s judgments were inaccurate at discriminating between well-learned (and less-well-learned) texts, whereas another group’s judgments were quite accurate. The accuracy of the latter group’s judgments was boosted by having the students generate keywords that captured the essence of each text prior to making metacomprehension judgments. We discuss this technique in greater detail below. After the initial study–judgment–test phase, participants selected any texts they wanted to restudy and restudied them, and then they answered six new criterion questions about each text. Performance from the initial test to the criterion test did not improve for the group with poor judgment accuracy, whereas test performance improved considerably for the group that made relatively accurate judgments. Critically, these gains in performance were linked to the accuracy of their judgments, which could be used to identify the texts that really needed to be restudied.

Unfortunately, people’s judgment accuracy is often poor, which we highlight in our brief history of metacomprehension research. After this review, we describe recent attempts to improve accuracy, which has led to optimism and new directions for future research.

BRIEF HISTORY OF METACOMPREHENSION RESEARCH

Seminal research on metacomprehension judgments was published in the mid 1980s and provided the most common methods

Address correspondence to John Dunlosky, Psychology Department, Kent State University, Kent, Ohio, 44242; e-mail: jdunlosk@kent.edu.

for assessing judgment accuracy (Glenberg & Epstein, 1985). In general, participants read multiple texts on a variety of topics (e.g., intelligence, weather, the French Revolution), which typically run at least 200 words in length and in some cases are as long as 1,000 words. After reading each text, participants are asked to make a single judgment about how well they will perform on a criterion test that covers the material covered in that text. The judgment often ranges from 0 (“I will not answer any questions correctly”) to 100 (“I’ll answer 100% of the questions correctly”).

Judgment accuracy is measured by the correspondence between each person’s judgments and his or her test performance. Two kinds of accuracy have been explored—*relative accuracy* and *absolute accuracy*. Relative accuracy is the degree to which a person’s judgments correlate with his or her own test performance across texts. The correlation ranges from -1 to $+1$, with correlations near or below 0 indicating poor accuracy. Ideally, people’s relative accuracy would be close to $+1.0$, which would indicate that they are excellent at discriminating between texts they have learned well and those they have not. Absolute accuracy pertains to whether a participant’s judgments are over or underconfident; for instance, if a person judges that he has learned all the content of a text yet fails the criterion test, then he is quite overconfident. Thus, for absolute accuracy, we are interested in whether the absolute value of people’s judgments match the absolute level of their test performance across texts, whereas for relative accuracy, we are interested in whether people’s judgments discriminate between their test performance for one test relative to another. Note that relative and absolute accuracy assess different aspects of accuracy, so that relative accuracy may be high when absolute accuracy is quite low. For instance, a person may judge that she has learned on average 80% of the material across 10 paragraphs. If her test performance across those texts is only 60%, then the absolute value of her judgments exceeds the absolute level of her test performance; in this case, she is overconfident. Even so, her relative accuracy could be perfect if her judgments increased or decreased with test performance—that is, if her test performance was better on paragraphs that were given higher judgments. Relative accuracy has been the mainstay of metacomprehension research and hence will be highlighted here.

Over two decades of research have demonstrated that individuals of all ages usually achieve only low levels of relative accuracy. In a review of research that used the standard method described earlier, Maki (1998) noted that, averaging across 25 studies in her own laboratory, relative accuracy was only .27. The same mean accuracy was obtained in published research from Dunlosky’s laboratory, across 36 different conditions. We believe two factors are largely responsible for these low levels of accuracy. The first is that people will have difficulties judging their learning and comprehension of texts if they have minimal understanding of the text content in the first place. Moreover, by having people make only a single judgment about how well they have learned each text, the standard method does not allow

people to indicate what they have (versus have not) learned *within* each text. Identifying these factors as potential culprits has led to the development of new techniques for improving metacomprehension accuracy.

IMPROVING METACOMPREHENSION ACCURACY

Approaches Using the Standard Method

Given the importance of accurate metacomprehension to efficient regulation of learning, we and others have been searching for techniques that will substantially improve people’s judgment accuracy. Our own approach has been guided by integrating theories of text comprehension with basic theories of monitoring. A major implication of this integration was that judgment accuracy would benefit from techniques that help people to monitor information indicative of deeper understanding of the text materials (for technical details, see Rawson, Dunlosky, & Thiede, 2000). For instance, techniques that encourage people to understand the general meaning of each passage (as opposed to just interpreting individual sentences within it) were expected to benefit accuracy. To evaluate this possibility, we used the standard method described above and had college students read various texts either once (as was the norm in previous research) or twice, because rereading encourages people to process a text for deeper understanding. As expected, relative accuracy was greater after rereading (.57) than after a single reading (.24; Rawson et al., 2000, experiment 1).

Similarly, Thiede and his colleagues have pursued another technique—summarization—that was expected to improve accuracy, partly because summarizing a text can encourage people to monitor how well they understand its meaning. In two experiments, Thiede and Anderson (2003) had college students read multiple texts covering a variety of topics. After all texts were read, one group made a global judgment for each text, whereas those in another group first summarized each text before judging their learning. Across two experiments, judgment accuracy was significantly greater when students summarized text content (correlations higher than .55) than when they did not (correlations less than .30), although the effect arose only when summarization was delayed after reading. Moreover, Thiede et al. (2003) investigated whether full-blown summarization was needed to reap its benefits. Instead of summarization, they asked students to generate five key terms that best captured the essence of each text. Simply generating five key terms boosted accuracy, but only when keyword generation was delayed after reading. This delayed-keyword effect was recently replicated by Thiede, Dunlosky, Griffin, and Wiley (2005), who argued that delayed keyword generation improved accuracy because it forced readers to evaluate their general understanding of the texts.

Of course, some of these manipulations (e.g., rereading) also improved test performance, so their influence on judgment accuracy may partly be due to improvements in test performance

and not due to improvements in monitoring skill per se (for arguments against this possibility, see Rawson et al., 2000). Note, however, that such concomitant improvements in both judgment accuracy and test performance would be welcome in applied settings; and as important, some manipulations (e.g., keyword generation) did not influence test performance yet still boosted judgment accuracy, which provides strong evidence that monitoring skill can be improved independently of test performance.

Although these advances are quite encouraging, the levels of accuracy typically obtained even under the most favorable conditions are not consistently high. For instance, from the published articles described above, we surveyed those conditions that successfully enhanced accuracy. Across 12 conditions, the mean level of accuracy was only $+ .56$, which is still far from perfect. Thus, although this research demonstrates that accuracy can be successfully improved, we decided to look beyond the standard method in hopes of developing new techniques that will support even higher levels of metacomprehension accuracy.

A New Method for Investigating Metacomprehension

One technique arose from an intuition about why the standard method itself may constrain people's judgment accuracy. Consider reading a 400-word passage from a textbook that included definitions of various key concepts that are relevant to the central topic. For instance, you may be reading a passage from a textbook on research methodology that includes basic definitions for various measurement scales, such as ordinal, nominal, and ratio scales. Let's also assume you are quite good at evaluating how well you have learned the definitions of each of the scales. After reading this passage, you would then need to make a single judgment that represented how well you learned all the text content, which we call a *global judgment*. It would seem quite remarkable if you were able to translate your accurate monitoring of learning for particular concepts within the passage into a single value. That is, how would your global judgment accurately represent that you knew you learned some definitions quite well, others less well, and still others not at all? Moreover, criterion tests often involve testing people's learning of specific concepts within a passage (e.g., "What is the definition of an ordinal scale?"). Thus, inherent in the standard method is a mismatch between the grain size of the judgment (which is global) and the grain size of the critical information that is subsequently tested (which concerns specific concepts).

Based on this rationale, we expected that if students evaluated their learning at a grain size that better matched the subsequent test, the accuracy of their judgments would be excellent. In our first attempt to evaluate this expectation, we used a twist on the standard method. College students read passages from actual textbooks that each included four critical definitions (e.g., definitions of four measurement scales). After reading a passage, they made a global judgment, and the new twist was that they

were also asked to make four *term-specific judgments*, one for each definition. For instance, they would be asked, "How well will you be able to recall the definition of *ordinal measurement* on the upcoming test?" The criterion test was recall of the key definitions, so note that the grain size of the term-specific judgments and the grain size of the criterion tests were identical. Initially, we believed this particular experiment was trivial because the anticipated outcome was obvious: The relative accuracy of the term-specific judgments should be nearly perfect, because to make them, students merely need to attempt retrieval of each key term and use the outcome of this retrieval attempt as a basis for the judgments. To our surprise, however, relative accuracy was not reliably greater for term-specific judgments ($+ .42$) than it was for global judgments ($+ .40$; from Dunlosky, Rawson, & McDonald, 2002).

Why might the relative accuracy of term-specific judgments, contrary to expectations, be so low? When one considers the possible processes that may influence term-specific judgments, two difficulties arise with respect to the rationale described above. First, we assumed that when students were making term-specific judgments, they would attempt a full-blown retrieval of the sought-after target and use it to inform their judgment. Second, we also assumed that, even if students were attempting to retrieve each sought-after definition, they would be able to accurately evaluate the quality of their recall.

In follow-up experiments, both of these assumptions were shown to be invalid. In one experiment, immediately before making a term-specific judgment, one group of students attempted recall of the definition by typing it into the computer, whereas another group of students did not. Relative accuracy was greater when students attempted to recall definitions prior to making their judgments ($+ .73$) than when they did not attempt prejudgment recall ($+ .57$; Dunlosky, Rawson, & Middleton, 2005).

Even though prejudgment recall did boost accuracy, there was also room for improvement. Participants' difficulties were revealed through further analysis of prejudgment recall, which showed that they did not always accurately evaluate the quality of their recall. In particular, participants often believed they knew a correct answer when they had recalled an entirely incorrect one. These results are presented in the left panel of Figure 1, where the mean judgment is plotted as a function of the outcome of prejudgment recall (Dunlosky et al., 2005, Experiment 2). Most relevant here, term-specific judgments showed overconfidence for commission errors, which is consistent with the conclusion that at least some students believed they knew an answer whenever something came to mind, regardless of its quality (Koriat, 1993). In a follow-up study, we attempted to improve the absolute accuracy of the term-specific judgments by providing feedback as individuals evaluated their learning (Rawson & Dunlosky, in press). Namely, participants first attempted to recall a definition, and then they were shown their responses along with the original definition (the feedback) when

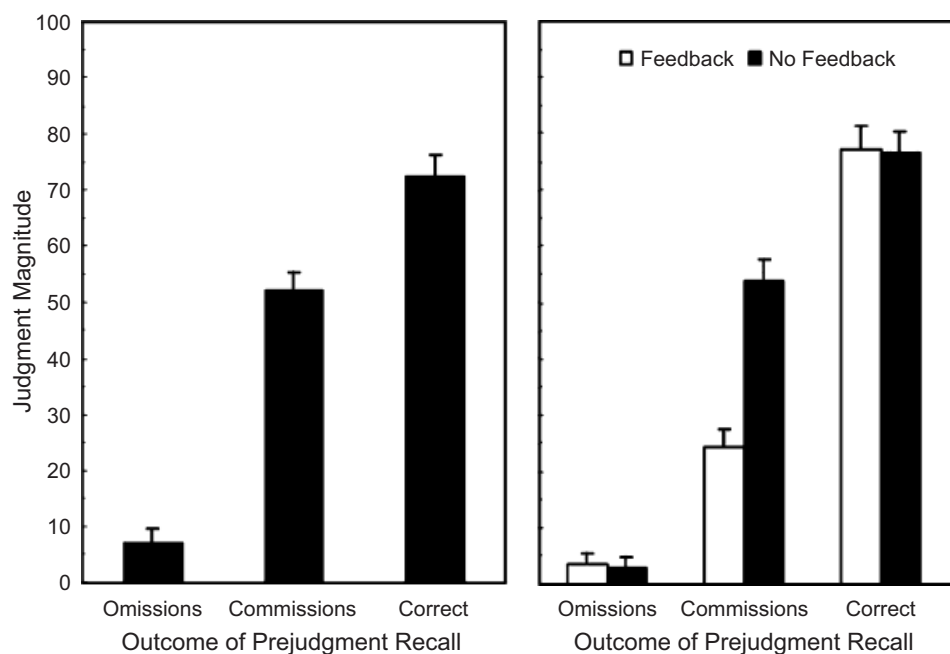


Fig. 1. Mean judgment as a function of the outcome of prejudgment recall. Omissions = nothing recalled; commissions = recalled response was entirely incorrect; correct = recalled response was entirely correct. Error bars are standard errors of each mean.

they made their judgments. As shown in the right panel of Figure 1, simply presenting the definition reliably reduced (although it did not eliminate) overconfidence for responses that were entirely incorrect. Moreover, students' term-specific judgments showed excellent relative accuracy (+.92) at discriminating responses that were correctly recalled from those that were not correctly recalled during prejudgment recall. These results are quite encouraging and provide optimism that new methods can be developed to improve students' metacomprehension.

CONCLUSIONS AND FUTURE DIRECTIONS

The implications from research on metacomprehension are both sobering and promising. Sobering, because research has repeatedly shown that global judgments of text learning and comprehension are not consistently (or impressively) accurate. Given that people may often want to evaluate their general learning or comprehension of texts, researchers have explored how to boost the accuracy of global judgments. One approach that has shown initial success involves encouraging people to read texts for a deeper level of understanding, such as by having them reread or summarize texts. Unfortunately, even these techniques do not consistently produce high levels of accuracy, so discovering what constrains the accuracy of global judgments and how to improve it remains an important aim for future research.

The implications are also promising, because term-specific judgments appear to produce robust and high levels of accuracy. Even here, however, a caveat is in order, which is well illustrated

by students preparing for an in-class examination. Students may often read a chapter of a text and then use the end-of-the-chapter terms (e.g., "Define *ordinal measurement*") to evaluate their learning, which is functionally similar to a term-specific judgment. We suspect that some students do not approach this activity properly and merely evaluate their learning based on their familiarity with the terms. Doing so would induce them to be overconfident in their knowledge and subsequently surprised at their failure—a circumstance that is much too common in the classroom. One implication here is that people should be coached on how to use these monitoring aids, such as by overtly attempting to recall answers and then by diligently checking them against appropriate feedback.

Even though term-specific judgments can support high levels of judgment accuracy, before they can be used widely to improve text learning, many questions will need to be answered. For instance, can term-specific judgments be adapted to help people monitor their comprehension (and not just learning) of text materials? Will even more specific judgments (e.g., at the level of individual idea units in each concept) serve to further reduce overconfidence and support even higher levels of accuracy? And, can individuals of all ages—from a fifth grader learning about light and sound in class to an older adult training for a new job—make these judgments accurately and use them to efficiently regulate their learning? By pursuing these questions in future research, we suspect that much progress will be made toward discovering techniques that consistently support excellent metacomprehension.

Recommended Reading

- Dunlosky, J., Rawson, K.A., & Middleton, E. (2005). (See References)
- Maki, R.H., & McGuire, M.J. (2002). Metacognition for text: Findings and implications for education. In T.J. Perfect & B.L. Schwartz (Eds.), *Applied Metacognition*. (pp. 39–67). New York: Cambridge University Press.
- Weaver, C. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 214–222.
-

Acknowledgments—Many thanks to Keith Thiede and the RADlab group for comments on a draft of this article. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H050038 to Kent State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

- Dunlosky, J., Hertzog, C., Kennedy, M., & Thiede, K. (2005). The self-monitoring approach for effective learning. *International Journal of Cognitive Technology*, *10*, 4–11.
- Dunlosky, J., Rawson, K., & McDonald, S. (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* (pp. 68–92). Cambridge, England: Cambridge University Press.
- Dunlosky, J., Rawson, K.A., & Middleton, E. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language*, *52*, 551–565.
- Glenberg, A.M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 702–718.
- Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*, 609–639.
- Maki, R.H. (1998). Test predictions over text material. In D.J. Hacker, J. Dunlosky, & A.C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–144). Mahwah, NJ: Erlbaum.
- Rawson, K., & Dunlosky, J. (in press). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*.
- Rawson, K.A., Dunlosky, J., & Thiede, K.W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, *28*, 1004–1010.
- Thiede, K.W., & Anderson, M.C.M. (2003). Summarizing can improve meta-comprehension accuracy. *Contemporary Educational Psychology*, *28*, 129–160.
- Thiede, K.W., Anderson, M.C.M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of text. *Journal of Educational Psychology*, *95*, 66–73.
- Thiede, K.W., Dunlosky, J., Griffin, T., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1267–1280.