

Using Standards to Improve Middle School Students' Accuracy at Evaluating the Quality of Their Recall

Amanda R. Lipko

The College at Brockport State University of New York

John Dunlosky, Marissa K. Hartwig,
Katherine A. Rawson, Karen Swan, and Dale Cook
Kent State University

When recalling key term definitions from class materials, students may recall entirely incorrect definitions, yet will often claim that these commission errors are entirely correct; that is, they are overconfident in the quality of their recall responses. We investigated whether this overconfidence could be reduced by providing various standards to middle school students as they evaluated their recall responses. Students studied key term definitions, attempted to recall each one, and then were asked to score the quality of their recall. In Experiment 1, they evaluated their recall responses by rating each response as fully correct, partially correct, or incorrect. Most important, as they evaluated a particular response, it was presented either alone (i.e., without a standard) or with the correct definition present. Providing this full-definition standard reduced overconfidence in commission errors: Students assigned full or partial credit to 73% of their commission errors when they received no standard, whereas they assigned credit to only 44% of these errors when receiving the full-definition standard. In Experiment 2, a new standard was introduced: Idea units from each definition were presented, and students indicated whether each idea unit was in their response. After making these idea-unit judgments, the students then evaluated the quality of their entire response. Idea-unit standards further reduced overconfidence. Thus, although middle school students are overconfident in evaluating the quality of their recall responses, using standards substantially reduces this overconfidence and promises to improve the efficacy of their self-regulated learning.

Keywords: judgment accuracy, metacomprehension, metamemory, student learning

In most school settings, students are expected to learn a large amount of material across a variety of content areas. This expectation is especially true for younger students, given the number of standardized tests administered in elementary and middle schools. Thus, efficiently learning a great deal of class materials is a major goal for students. One method that can help students achieve this goal is to accurately evaluate the quality of their learning while studying course materials (Dunlosky, Hertzog, Kennedy, & Thiede, 2005; Winne, 2004). The reason why accurate evaluations

can improve the efficiency of learning arises directly from models of self-regulation that emphasize the interactive nature of monitoring and control processes (e.g., Nelson & Narens, 1990; Winne & Hadwin, 1998). In Winne and Hadwin's (1998) model, students set goals for achievement, monitor ongoing progress while learning, and evaluate whether their goals have been met (for a review, see Greene & Azevedo, 2007). If their monitoring indicates that their goals have been met, then students will move on. If monitoring indicates that their goals have not been met, students may engage any number of control processes to achieve their goals. These control processes may include making further effort using the same strategy to obtain the goal, choosing another strategy to obtain it, or even deciding to change the sought-after goal. Students' evaluation of their progress may also influence regulation of learning when subsequent study is guided by an automated tutor or other agent, such as a parent, tutor, or teacher, because these agents may use those evaluations to help guide the students' learning. Either way, accurate monitoring is needed to achieve effective control.

In the case of learning concepts for classroom exams, if students are capable of identifying concepts that they have not learned well, their learning could be effectively guided by focusing their efforts on those concepts. They may decide to spend extra time studying them or may seek help from a peer or teacher. If they judge that a concept has been learned when it has not, then control will be less effective. In this case, the student may fail to restudy the unlearned concept; similarly, if an automated tutor is controlling restudy on the basis of the student's evaluation of learning, the tutor may inadvertently drop the unlearned concept from study. Accordingly,

Amanda R. Lipko, The College at Brockport State University of New York; John Dunlosky, Marissa K. Hartwig, and Katherine A. Rawson, Dept of Psychology, Kent State University; Karen Swan and Dale Cook, Research Center for Educational Technology, Kent State University.

Many thanks to the RADlab for discussion and feedback on these projects and to Pat Mazzer, Frank Seman, Thomas McNeal, and Mark van't Hoofft for their valuable assistance collecting data from middle school students in the AT&T Classroom at the Research Center for Technology (www.rcet.org) at Kent State University. We thank Melissa Bishop and Cristin Clewell for assistance in scoring participants' recall protocols.

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H050038 and Grant R305A080316 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. It was also partially supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior Collaborative Award.

Correspondence concerning this article should be addressed to John Dunlosky, Kent State University, Psychology Department; Kent, OH 44242. E-mail: jdunlosk@kent.edu

discovering techniques to improve students' ability to accurately evaluate their learning promises to improve their success (Nietfeld, Cao, & Osborne, 2006; Thiede, 1999).

More specifically, imagine students studying a list of key term definitions from a unit on genetics. They would study all of the to-be-learned definitions (see Figure 1A). Then, to judge how well they know the definitions, they would first attempt to recall a previously studied definition. In the example shown in Figure 1B, the student's recall response for "What is homozygous?" is "something from your parents that ends up in your genes." We refer to this recall attempt as *prejudgment recall* because it occurs immediately before a student judges a recall response (Nelson, Narens, & Dunlosky, 2004). After attempting prejudgment recall for a particular definition, the student would make a *self-score judgment*, which involves rating the quality of the prejudgment recall (for definitions of key terms used throughout this article, see Table 1). In Figure 1C, the student's self-score judgment for the response "What is *homozygous*?" was "full credit." These self-score judgments, which are the focus of the present research, will be most beneficial for guiding study if students *accurately* judge the quality of their recall. In particular, if students can accurately identify their recall errors, they can allocate additional time to studying those definitions. However, if they make errors that they do not identify, they may cease studying definitions before they have been mastered. In the example in Figure 1, the student's self-score judgment was overconfident; that is, the student's response for "What is *homozygous*?" is entirely incorrect, which is called a *commission*

Table 1
Definition of Terminology Used to Refer to Key Components of the Experimental Method

Term	Definition
Prejudgment recall	Attempted recall of key term definition made immediately prior to making judgments (Figure 1B)
Self-score judgment	Participant's judgment of the quality of prejudgment recall response for a given definition (Figure 1C)
Full-definition standard	Providing the entire correct definition as a standard when making a self-score judgment
Idea-unit standard	Providing the individual idea units in the correct answer and the participant's idea-unit judgments when making a self-score judgment
Idea-unit judgment	Participant's judgment about whether an idea unit from the correct response of a definition is in the prejudgment recall response (Figure 3)

error. In this case (and with all commission errors in the present experiments), a teacher would give the student no credit. However, the student evaluated the response and judged it as entirely correct, which demonstrates overconfidence.

The general idea here is simply that overconfidence can lead to suboptimal learning, whereas accurate judgments will support more effective learning (Thomas & McDaniel, 2007; Winne, 2004). The former is especially problematic given that students do not choose to restudy items that they believe they already know (e.g., Metcalfe & Kornell, 2005); so, if students incorrectly judge that a commission error is entirely correct, they will be unlikely to fix that error through subsequent restudy. Unfortunately, like in the example presented in Figure 1, students are often overconfident in their learning of text materials. For example, Rawson and Dunlosky (2007) asked college students to read several expository texts taken from textbooks that contained key term definitions (e.g., a text on psychological measurement containing definitions of four measurement scales, including "Ordinal measurement: Ranking on a continuum where the differences between consecutive values are not necessarily equal"). After reading each text, participants attempted prejudgment recall for the key term definitions (as in Figure 1B). Immediately after they typed a given response, they were asked (similar to Figure 1C), "If the correctness of the definition you just wrote was being graded, do you think it would receive: no credit, partial credit, or full credit?" To measure the accuracy of these self-score judgments, students' prejudgment recall responses (as in Figure 1B) were later scored by separating them into several categories, including entirely correct, partially correct, and commission errors. Next, the judgments (as in Figure 1C) were analyzed as a function of each kind of recall response. Most important, when recall responses were entirely incorrect, college students' judgments demonstrated overconfidence. In particular, they erroneously awarded partial or full credit to 83% of these commission errors. Such overconfidence is troubling, especially because commission errors are common when people learn new definitions (Dunlosky, Rawson, & Middleton, 2005; Kikas, 1998; Rawson & Dunlosky, 2007).

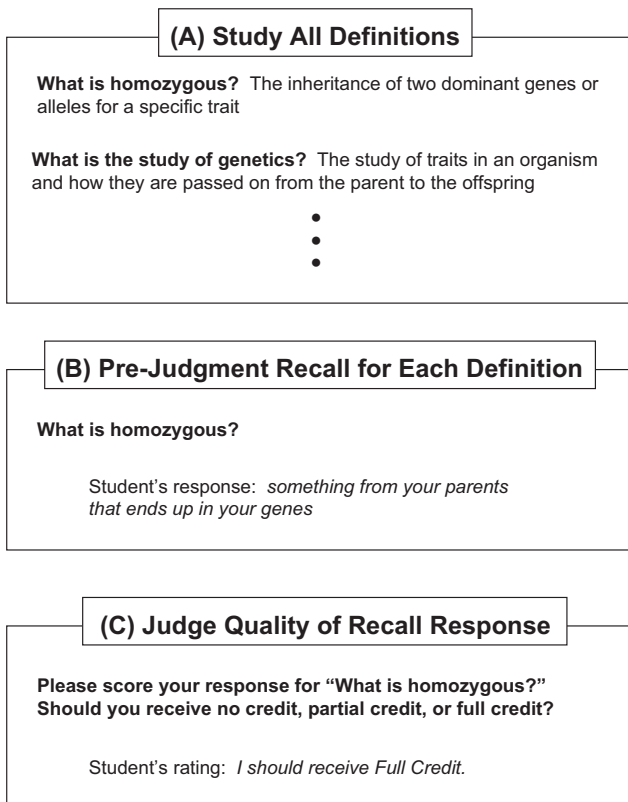


Figure 1. Illustration of how students can study and evaluate their learning of key definitions from classroom materials. See text for details.

In the present research, we established that middle school students also exhibit marked overconfidence when making self-score judgments for recently learned key term definitions, and we explored techniques to improve the accuracy of these judgments. Similar to the procedure illustrated in Figure 1, we had students study definitions, engage in prejudgment recall for the correct answers, and then make self-score judgments in which they scored the quality of their recall. We were most concerned with the degree to which providing various standards of evaluation (such as providing the correct definition while students evaluated their responses) would reduce their overconfidence for commission errors. One research question was of greatest interest: Will the accuracy of middle school students' judgments benefit from providing various standards as they evaluate the quality of their recall responses?

We first describe theory of overconfidence that is relevant to improving the accuracy of students' self-score judgments. We then discuss previous research relevant to the influence of providing full-definition standards on college students' overconfidence in their commission errors. Finally, we justify our current focus on middle school students, as well as explain secondary analyses aimed at evaluating the potential quality of control decisions based on self-score judgments.

Why Are Students' Judgments Overconfident?

Why might students give themselves credit for an incorrect answer? And more important for the present investigation, how might such overconfidence be reduced? To address these questions, our research has been guided by theoretical frameworks that assume people do have some inherent difficulties in accurately judging their learning, but also that overconfidence can be reduced by providing appropriate environmental cues. *Cue-utilization frameworks* state that people's judgments of learning for text (also known as *metacomprehension judgments*) are inferential in nature and are based on any number of cues that people view as relevant to how well text materials have been learned (e.g., Dunlosky, Rawson, & Hacker, 2002; Rawson, Dunlosky, & Thiede, 2000; and see Koriat, 1993, 1997). Consistent with such frameworks, people apparently use multiple cues to make metacomprehension judgments, such as perceived familiarity with the domain covered by texts (Maki, 1998), momentary access to the text content (J. Baker & Dunlosky, 2006; Morris, 1990), the propositional density of a text (Miles & Stine-Morrow, 2004), and the ease of processing text (Rawson & Dunlosky, 2002). Unfortunately, many of these cues are not diagnostic of how well one has learned a text, and such cues can produce inaccuracy and overconfidence.

Most relevant here is Koriat's (1993) *accessibility hypothesis*, which is a specific instantiation of cue-utilization frameworks. According to this hypothesis, when people are judging whether an answer is correct, one cue is how much information they accessed about the to-be-judged material, regardless of whether the information is correct. For making self-score judgments of prejudgment recall, the accessibility hypothesis predicts that students will give themselves credit if they access any information about a definition whatsoever, regardless of whether the information accessed is correct or incorrect. That is, even if only incorrect information is accessed (as with commission errors), accessing anything will lead students to believe they are partially or entirely correct.

To counteract such overconfidence, Rawson and Dunlosky (2007) argued that students must compare their answer with an appropriate standard. In particular, they proposed that students were basing their self-score judgments on the amount of information accessed and, hence, were not fully scrutinizing the quality of the information that was accessed (cf. L. Baker, 1984; Dunlosky, Rawson et al., 2005). Thus, if students are given the correct definition as an external standard to evaluate the quality of their recall, it may help them identify which ideas in their recall response are incorrect. Within the cue-utilization framework, the standard of evaluation provides an external cue that students can use to better judge the quality of their prejudgment recall.

The Influence of Standards on Reducing Overconfidence

To evaluate this prediction, Rawson and Dunlosky (2007) had college students study short texts containing key term definitions, attempt prejudgment recall for each definition when cued with the term, and then make self-score judgments for their prejudgment recall responses. For some participants, the correct definition was used as a standard of evaluation (called a *full-definition* standard; see Table 1), which was presented as they made self-score judgments for each of their responses. As expected, this standard did significantly reduce (although did not eliminate) students' overconfidence.

In the present experiments, a major goal was to evaluate whether children's self-score judgments would also benefit from using various standards of evaluation. Given that metacognition develops through grade school (Schneider & Pressley, 1997), the question arises as to whether children will also be competent at evaluating their recall of key term definitions when they receive standards. Although an extensive literature on children's memory monitoring exists, many of these studies relied on traditional metacognitive methods with simple items, such as predicting memory performance for paired associates (for a review, see Schneider & Lockl, 2008). Most important, no previous study has investigated the degree to which children can judge the quality of their learning of key term definitions.

Accordingly, as a first step to filling this gap in the literature, we sought to investigate middle school students' skill at judging their learning of key term definitions. We chose to focus on middle school students (vs. even younger students) for several reasons. First, middle school students appear to be at least partially deficient at spontaneously detecting errors in text materials as compared with students in higher grades (Hacker, 1997). Thus, we expected their judgments to be overconfident. Second, assuming that middle school students are overconfident without a standard, one might expect that standards would help them evaluate the quality of their recall. For instance, middle school students' skill at detecting errors in text improves when they are instructed about the kinds of error that they should be searching for when reading texts (L. Baker, 1984; Hacker, 1997; Zabrocky & Moore, 1989). Given this evidence, one prediction is that full-definition standards will reduce middle school students' overconfidence in commission errors. By contrast, preschoolers and first graders appear overconfident even when given explicit feedback about their failures (e.g., Lipko, Dunlosky, & Merriman, 2009). Thus, we chose middle school students because the accuracy of their judgments may also be improved by using standards, whereas even younger students'

judgments may not. If middle school students do benefit, then it seems reasonable that the judgments of older students (high school students) would also benefit, and future research could explore possible developmental trends in even younger children.

Overview of Experiments

In Experiment 1, middle school students studied definitions about genetics and literary nonfiction. After attempting prejudgment recall, they evaluated the quality of their recall either with or without a full-definition standard. Given the evidence from college students (Rawson & Dunlosky, 2007), we expected that self-score judgments would be lower for commission errors when made in the presence of a standard than without one. In Experiment 2, we compared the influence of a full-definition standard against that of an idea-unit standard, which is described in more detail below.

We also present analyses relevant to the potential quality of control decisions based on the students' judgments. Research on control effectiveness has largely examined the quality of control decisions in two situations: (a) self-regulated study, where students decide which items to restudy (e.g., Ariel, Dunlosky, & Bailey, 2009; Son & Metcalfe, 2000), and (b) experimenter-paced study, where an algorithm is used to decide which items will be restudied (e.g., Nelson, Dunlosky, Graf, & Narens, 1994). In the present case, we focus on the quality of control decisions made using an algorithm, which simulates the decisions that an automated tutor may make about whether items could be dropped from study on the basis of the students' evaluations. For instance, one decision would be to drop any definition from further study that receives a self-score of "entirely correct." If students' self-score judgments are used in this manner, would the tutor drop items that were incorrectly recalled and fail to drop items that were correct? We describe analyses that answer this question below.

Experiment 1

Method

Participants. Eighteen students from an eighth-grade class participated (mean age = 13.0 years, $SD = 4.2$ months; 11 girls). Students were recruited from a northeast Ohio middle school by the Research Center for Educational Technology at Kent State University. At the time of data collection, their class was being held at the AT&T Classroom at Kent State University, where they were receiving daily instruction for several weeks.

Design and materials. The most critical variable was the kind of standard (none vs. full-definition standard), which was a within-subject manipulation. Two kinds of definitions (genetics definitions and literary nonfiction definitions) were also used (within-subject manipulation) to evaluate whether any effects would generalize across materials (for examples, see the Appendix). For each kind of material, 20 definitions were separated into two definition sets of 10 definitions (i.e., Sets A and B). All definitions used in these experiments were developed with a middle school teacher's assistance, so that the definitions represented the kinds of material that students were expected to learn in their classes.

To test students with all four definition sets, we used four sessions, each separated by 1 week. During a given session, participants studied one material set of 10 definitions (either ge-

netics or literary nonfiction) and made judgments with either no standard or with a full-definition standard. Note that multiple sessions were used because we had only 30 min to interact with students during any one class period; thus, we separated the experiment into meaningful units (e.g., study one kind of material and use one kind of standard) that would fit within this time frame.

Genetics definitions were presented in the first two sessions, and literary nonfiction definitions were presented in the second two sessions. It is important to point out that for the two sessions with each kind of material, we counterbalanced the order of standard (no standard in the first session vs. full-definition standard in the first session) and the order of material set (Set A in first session vs. Set B in first session). Students were randomly assigned to the counterbalancing groups to ensure approximately equal numbers of students per group. In this and the following experiment, we initially conducted analyses that included counterbalancing orders as factors. Their effects were never statistically significant, so we collapsed across all order factors and do not discuss them further. Computers presented all materials and recorded all responses.

Procedure. Participants worked individually at computers. They were encouraged to take their time, to read the instructions carefully, and to work quietly. They began each session by reading detailed instructions that were presented on the computer screen. Following the instructions, the key term definitions were presented one at a time for self-paced study. Thus, as shown in Figure 1A, a participant would first study each definition (i.e., the question stem, along with the answer) at his or her own pace until all the definitions had been presented for study. Participants clicked a button on the screen to indicate when they were finished studying each definition. Order of presentation was randomized anew for each participant.

Immediately after studying all the definitions, the participant was presented with a key term (e.g., "What is *homozygous*?") and attempted to recall its correct definition (i.e., a prejudgment recall attempt; see Figure 1B). Participants were instructed to try their best to generate the answer or as much of the answer as possible for each key term. They typed their answers into a text field on the screen. After this prejudgment recall attempt, they then judged the quality of their prejudgment recall for that key term. In the full-definition standard condition, immediately after attempting prejudgment recall for a key term, participants were shown the correct definition along with their response and were then asked to make a self-score judgment using the options *no credit*, *partial credit*, or *full credit*. They were instructed to make this judgment by comparing their response with the correct answer. In the no-standard condition, participants were shown only their response when making a self-score judgment.

The prejudgment recall attempts and all the judgments were paced by the participants. After completing the prejudgment recall and judgment trial for a given key term definition, the next key term was presented for prejudgment recall and judgment. These recall judgment trials continued until they had attempted to recall and judge each item.

Results

In both experiments, all planned comparisons declared as significant had $p < .05$, one-tailed, and the alpha value for all post hoc comparisons was corrected using the Bonferroni adjustment.

Prejudgment recall. Before reporting our most critical outcomes involving the self-score judgments, we briefly present analysis of prejudgment recall. Each prejudgment recall response was scored as *entirely correct* (i.e., it included all the main ideas, either verbatim or correctly paraphrased), as *partially correct* (i.e., contained one or more of the correct ideas but not all of them), or as *incorrect* (i.e., contained none of the correct ideas). Correctly recalled responses were scored as 100 (for 100% correct), partially correct recall responses were scored as 50 (for half credit), and incorrectly recalled responses were scored as 0. Finally, two assistants independently scored recall responses from 10 participants, and the agreement of their scoring was high (94%); the remaining recall responses were scored by one assistant. As important, the assistants were blind with respect to participants' group assignment and to participants' self-score judgments for any key term definition.

Means across participants' prejudgment recall scores were as follows: for genetics definitions, 0.27 ($SE = 0.04$) for the full-definition standard condition and 0.28 ($SE = 0.04$) for the no-standard condition; for literary definitions, 0.40 ($SE = 0.04$) for the full-definition standard condition and 0.37 ($SE = 0.05$) for the no-standard condition. An analysis of variance (ANOVA) revealed only a significant effect of materials, with students performing slightly better on the literary definitions than on the genetics definitions, $F(1, 13) = 16.0$, $MSE = 0.011$.

As important, we also separated the prejudgment recall responses into one of four categories: *omission error* (i.e., the student provided no response during prejudgment recall), *commission error* (i.e., recall response was completely incorrect), *partially correct* (i.e., recall response contained at least one correct idea unit but not all of them), and *correct recall* (i.e., recall response contained all of the ideas from the correct answer). The percentages of each prejudgment recall response are presented in Table 2. Note that commission errors were relatively common.

Self-score judgments. To investigate how well students judged the quality of their prejudgment recall, we analyzed the self-score judgments as a function of the categories of prejudgment recall responses. For consistency with our previous research (e.g., Dunlosky, Rawson et al., 2005; Rawson & Dunlosky, 2007), we

first assigned a value of 0 to self-score judgments of no credit, 50 for partial credit, and 100 for full credit. We then computed the mean self-score judgment for responses within each of the four categories of prejudgment recall. Most important, for commission errors, mean self-score judgments greater than 0 indicate overconfidence. The values are reported in Figure 2, which presents the mean self-score judgment (y-axis) conditionalized on prejudgment recall category (x-axis). Given that omission errors had a median self-score value of 0, we omitted this uninformative category from the figure and analysis.

Our analytic plan was to focus first on planned comparisons involving self-score judgments for commission errors. We then conducted a series of post hoc comparisons to investigate the significance of trends involving other responses. We favored this analytic plan over conducting standard omnibus ANOVA for two reasons. First, missing values arose when a participant missed a class day (and hence did not provide data for a particular condition) or when a participant did not make a particular kind of recall response (e.g., no partially correct responses, and hence a self-score judgment would not be available for that kind of response). Given that all but a few participants were missing at least one value and all key factors were manipulated within participant, an omnibus ANOVA would exclude nearly all the participants from the analysis. Second, and more important, in research contexts where a priori predictions are available, planned comparisons to evaluate them are superior to conducting omnibus ANOVAs (Judd & McClelland, 1989).

As evident from inspection of Figure 2, the presence of a standard decreased self-score judgments for commission errors. This reduction in overconfidence was statistically significant both for the genetics definitions, $t(12) = 2.68$, Cohen's (1998) $d = 0.72$, and for the literary nonfiction definitions, $t(15) = 4.66$, $d = 1.20$. Nevertheless, for both material sets, mean self-score judgments with the standard were still significantly greater than 0, $t_s > 4.5$, which indicated that some overconfidence for commissions remained. We also computed the percentage of commission errors that received the various judgment ratings (full credit, partial credit, and no credit), which are presented in Table 3. Collapsed across material, students assigned full or partial credit to 44% of their commission errors in the full-definition condition, whereas they assigned credit to 73% of their commission errors in the no-standard condition. Thus, full-definition standards do reduce but do not eliminate middle school students' overconfidence in judging their commission errors.

As in previous research with college students (e.g., Dunlosky, Rawson et al., 2005), self-score judgments tended to increase with the amount of correctly recalled information; that is, they increased from commission errors to partially correct responses, and then again to correct responses. Such increases were significant for all comparisons, $p < .006$ ($\alpha = .006$ by Bonferroni adjustment), except (a) between partially correct and correct recall for genetics definitions both with the full-definition standard, $p = .18$, and without one, $p = .45$, and (b) between commission errors and partially correct recall for the literary nonfiction questions with no standard, $p = .03$. Even with these exceptions, it is apparent that middle school students do discriminate to some extent between incorrect and correct responses.

Quality of control decisions based on self-score judgments. Our focal questions concerned the accuracy of students' judgments, so we did not have them make decisions about how to

Table 2
Percentage of Prejudgment Recall Responses for Each Response Category

Material/standard	Omission	Commission	Partial	Correct
Experiment 1				
Genetics				
Full definition	17.1	42.9	26.4	13.6
No standard	16.4	42.9	26.4	14.3
Literary nonfiction				
Full definition	3.6	40.7	32.1	23.6
No standard	2.9	47.1	26.4	23.6
Experiment 2				
Idea unit	2.3	38.5	41.8	17.4
Full definition	1.0	33.1	46.5	19.5
No standard	2.5	33.3	42.6	21.6

Note. Materials in Experiment 2 included only literary nonfiction concepts.

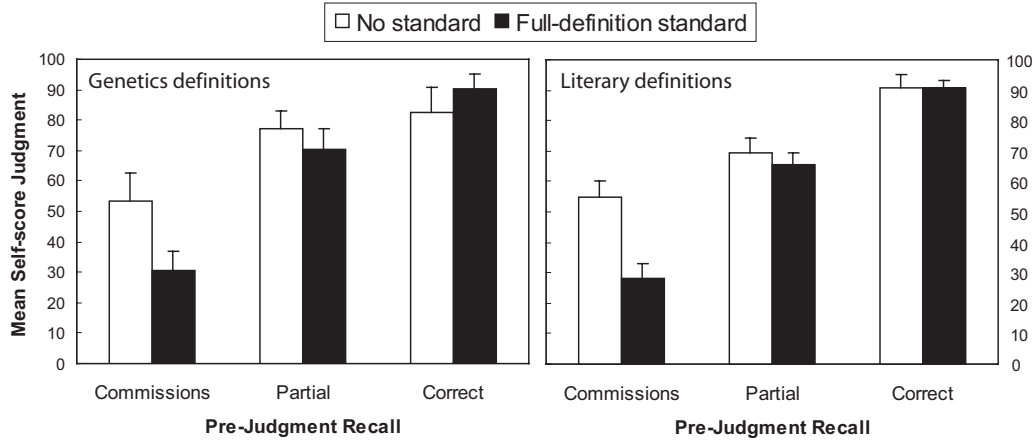


Figure 2. For Experiment 1, bars represent mean self-score judgment magnitudes for categories of objectively scored responses made during prejudgment recall. Left panel includes values for genetics definitions, and the right panel includes values for literary nonfiction definitions. Error bars are standard errors of the mean.

allocate restudy time across definitions. Nevertheless, students' judgments can be used by other agents (e.g., an automated tutor) to make allocation decisions, and the quality of these decisions can be evaluated in part by computing the likelihood that an item would be dropped from study given a particular judgment. In the present case, our specific question was, If a definition is dropped from study when students made a self-score judgment of "fully correct," would the automated tutor make better decisions when students made judgments with the full-definition standard than without it? So, for instance, would it drop fewer incorrectly recalled definitions (i.e., commission errors) and drop more correctly recalled ones?

In Table 4, we computed the proportion of definitions that would be dropped from study when students judged that they had

correctly recalled the definition. Given that analyzing all prejudgment recall responses (commission, partially correct, and correct) in an omnibus ANOVA would exclude most participants, we conducted separate 2 (standard) × 2 (material) ANOVAs for each type of recall response. Most important, for commission errors, the main effect of standard was significant, $F(1, 11) = 17.1, MSE = 0.32$, demonstrating that fewer definitions that students did not correctly recall would be dropped from study when judgments were made with a full-definition standard than without one. For partially correct and correct responses, no effects were significant. Thus, although this analysis was exploratory, it does suggest that

Table 3
Percentage of the Self-Score Judgments for Commission Errors

Material/standard	Self-score judgment					
	No credit		Partial credit		Full credit	
	M	SE	M	SE	M	SE
Experiment 1						
Genetics						
Full definition	54	9	31	7	15	5
No standard	35	10	24	5	41	10
Literary nonfiction						
Full definition	58	7	28	7	14	5
No standard	19	6	52	7	29	7
Experiment 2						
Idea unit	54	4	35	4	8	2
Full definition	44	5	43	5	16	4
No standard	15	3	55	6	33	6

Note. Commission errors were student responses during prejudgment recall that were objectively scored as entirely incorrect. Materials in Experiment 2 included only literary nonfiction concepts.

Table 4
Proportion of Definitions That Would Be Dropped From Study for Each Response Category

Material/standard	Commission		Partially correct		Correct	
	M	SE	M	SE	M	SE
Experiment 1						
Genetics						
Full definition	.15	.05	.47	.10	.80	.10
No standard	.41	.10	.58	.11	.73	.09
Literary nonfiction						
Full definition	.14	.05	.37	.08	.81	.05
No standard	.35	.07	.45	.09	.71	.10
Experiment 2						
Idea unit	.07	.03	.23	.07	.55	.07
Full definition	.23	.06	.59	.04	.96	.02
No standard	.32	.06	.59	.05	.84	.06

Note. For the full-definition and no-standard conditions, the proportion of dropouts include only those definitions in which students rated a recall response as entirely correct with the self-score judgments. For the idea-unit standard, the proportion of dropouts include only those definitions in which students indicated that all the idea units were present in a prejudgment recall response. Materials in Experiment 2 included only literary nonfiction concepts.

students' self-regulated learning would benefit from using full-definition standards while evaluating the quality of their recall responses.

Experiment 2

Experiment 1 demonstrated that providing a full-definition standard reduced the overconfidence of middle school students' self-score judgments for commission errors. Even so, students still erroneously awarded credit to their commission errors even when they received the standard. Why might students still be overconfident when they are given a standard to which they can compare their answer?

One possible answer is that, although helpful, a full-definition standard is underinformative because it does not indicate which aspects of the definition must be in students' answers for them to receive partial or full credit. Consider the example in Figure 1, which illustrates a student judging the quality of prejudgment recall without a standard. Now assume that the student received the full-definition standard; that is, the definition of *homozygous* was presented while the student made the self-score judgment. In this case, the fact that the prejudgment recall includes a word that is in the correct answer (i.e., *genes*) may lead the student to believe that the response should receive "partial credit," even though the key idea of "inherit two dominant genes" was not recalled. Put differently, although some word-level overlap exists between the correct definition and the student's prejudgment recall, it is evident that none of the idea units are included in it. Thus, with full-definition standards, students may not realize what ideas are required to obtain partial or full credit.

According to the cue-utilization framework, if cues are available that provide information about which ideas need to be in a prejudgment recall response to receive credit, then students could use them to recalibrate their self-score judgments. This possibility suggests that providing idea-unit standards will reduce students' overconfidence in their commission errors. An *idea-unit standard* involves parsing the definition into smaller idea units (see Table 1). For example, the definition of *homozygous* may be parsed into two idea units: (a) inheritance of two dominant genes, and (b) for a specific trait. Prior to making a self-score judgment, students are asked to identify whether each of the idea units was in their prejudgment recall. Idea-unit standards were expected to provide even more informative cues with which to judge the quality of prejudgment recall because now students are provided with the smallest units of conceptual information that are required to receive credit. We predicted that if middle school students make idea-unit judgments immediately before scoring their prejudgment recall, their overconfidence in commission errors would be reduced even further.

Although less relevant to our current focus on judgment accuracy, idea-unit standards may also support better control of restudy. As shown in Table 4, even with the full-definition standard, an automated tutor would have dropped about 15% of the students' commission errors on the basis of their self-score judgments. Perhaps if definitions are dropped from study when students indicate that all ideas are present in a recalled response, even fewer definitions that were incorrectly recalled would be dropped from restudy.

To evaluate these possibilities, the procedure in Experiment 2 was similar to the one used in Experiment 1, with one important addition. Middle school students received either no standards (control group), full-definition standards when making their self-score judgments (as in Experiment 1), or first indicated which idea units were present in their answer (idea-unit judgments) before making their self-score judgments. If idea-unit standards yield even greater improvements in accuracy and control, this standard may be preferred in applied settings.

Method

Participants. Students ($N = 103$) from four eighth-grade classes participated. None of the students participated in Experiment 1. Classes were recruited from a northeast Ohio middle school by the Research Center for Educational Technology at Kent State University. The students in each class participated at their school and were tested as a group at individual computers located in a computer lab in the school library.

Design, materials, and procedure. We used a 3 (standard: none, full definition, idea unit) \times 2 (definition set: A and B) \times 2 (order of definition set: Set A on first session or Set B on first session) mixed design. Within each classroom, participants were randomly assigned to the groups to ensure an approximately equal number of participants per group. The final number of participants in the three experimental groups were 33 (no standard; mean age = 13.8 years, $SD = 5.2$ months; 13 girls), 34 (full-definition standard; mean age = 13.9 years, $SD = 3.9$ months; 17 girls), and 36 (idea-unit standard; mean age = 13.8 years, $SD = 4.5$ months; 17 girls).

Procedures were similar to Experiment 1, except for the following changes. First, we used only literary nonfiction concepts; participants studied and judged their learning of eight concepts during the first session, and repeated the procedure with eight new concepts during the second session. Sessions were separated by 2 days. Second, for the idea-unit standard group, participants made idea-unit judgments prior to scoring each recall response. In particular, immediately after they attempted prejudgment recall for a key term (see Figure 1B), the participants were shown the main ideas in the correct answer, which were presented together as illustrated in Figure 3 for the key term *biography*. Participants were instructed to check the box next to each idea that they believed was present in their response. If none of the ideas were present, they were instructed to check the box labeled, "None of the ideas are present in my answer." Instructions stressed that only those idea units that were actually in their answer should be checked off as present, and they received some practice making the idea-unit judgments (with practice concepts and responses) that included feedback. After they finished identifying idea units, participants were shown their response along with the idea units and their idea-unit judgments and were asked to make a self-score judgment.

Results

Analyses were collapsed across the two sessions, and if a participant completed only one session ($n = 13$; e.g., due to being absent from class), then values from the single session were included.

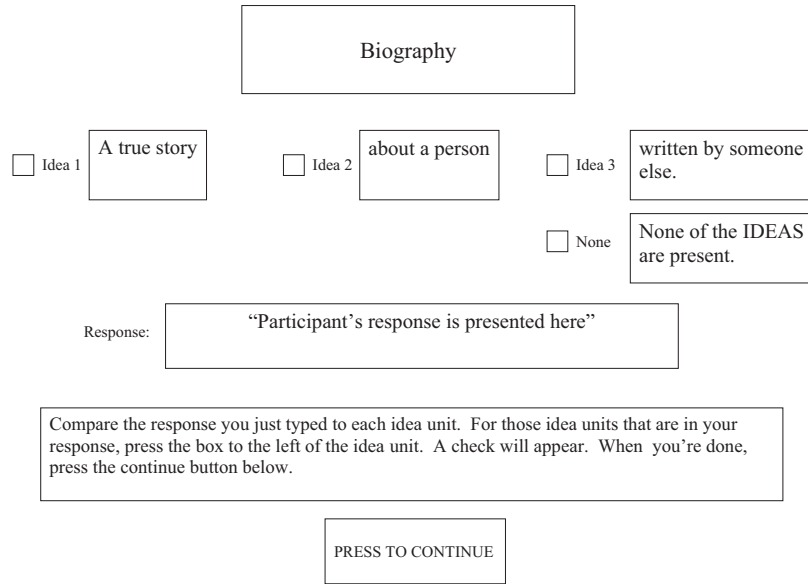


Figure 3. Screen format used to collect idea-unit judgments.

Prejudgment recall. Means across participants' prejudgment recall scores were as follows: 0.43 ($SE = 0.03$) for the no-standard group, 0.43 ($SE = 0.03$) for the full-definition group, and 0.38 ($SE = 0.03$) for the idea-unit group. The effect of standard was not significant, $F(2, 100) = 0.80$, $MSE = 0.03$. As in Experiment 1, commission errors were common (see Table 2).

Self-score judgments. As in Experiment 1, we first focus on planned comparisons involving self-score judgments for commission errors and then report post hoc comparisons to investigate the significance of trends involving other responses. Inspection of Figure 4 reveals two critical effects. First, self-score judgments for commission errors were markedly lower for students receiving the

full-definition standard than for those receiving no standard. Second, and more important, idea-unit judgments further reduced overconfidence in commission errors. Inferential analyses were consistent with these observations. For commission errors, we conducted planned comparisons to evaluate each standard group to the appropriate comparison group. Self-score judgments for commissions were lower for the full-definition group than for the no-standard group, $t(62) = 3.7$, $d = 0.93$, and self-score judgments were lower for the idea-unit group than for the full-definition group, $t(65) = 1.92$, $d = 0.47$. As in Experiment 1, we computed the percentage commission errors that received each of the judgment ratings (full credit, partial credit, and no credit). As illustrated in Table 3, participants assigned full credit to only 8% of their commission errors when making idea-unit judgments, yet assigned full credit to 16% of their commission errors when using the full-definition standards.

Finally, self-score judgments increased with the amount of correctly recalled information. Within each standard group, self-score judgments increased from commission errors to partially correct responses and increased again with correctly recalled responses. These increases were significant for all post hoc comparisons, $ps < .006$ ($\alpha = .006$ by Bonferroni adjustment), except for the difference between self-score judgments for partially and fully correct responses for the no-standard group, $p = .03$. Thus, middle school students show some ability to discriminate between responses that are correct versus incorrect.

Accuracy of the idea-unit judgments. To further explore students' skill at judging the quality of their responses, we also directly assessed the accuracy of their idea-unit judgments. In particular, we were interested in the situation in which a student claimed that his or her prejudgment recall included an idea unit from the correct response. When participants reported that an idea unit was in their response, was it actually there? Two values were most relevant here. *Correct identifications* were defined as the probability that a participant reported that an idea unit was present,

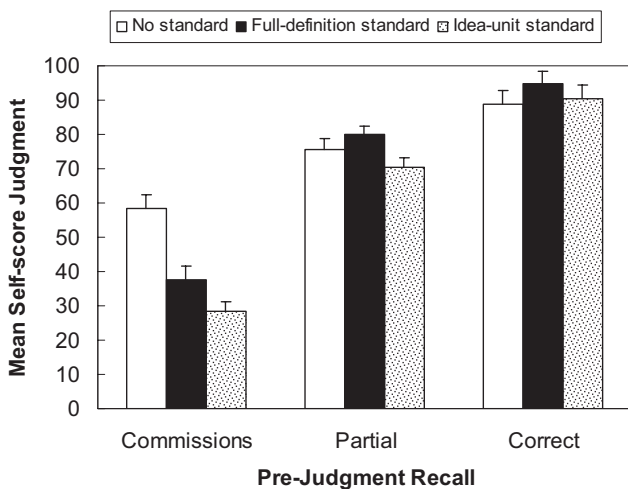


Figure 4. For Experiment 2, bars represent mean self-score judgment magnitudes for categories of objectively scored responses made during prejudgment recall. Materials were literary nonfiction definitions. Error bars are standard errors of the mean.

and the idea unit from the correct definition was actually present in the participant's prejudgment recall. *Incorrect identifications* were defined as the probability that a participant reported that an idea unit from the correct definition was present in prejudgment recall, but that it was actually absent. Ideally, students' correct identification rate would be 1.0, and their incorrect identification rate would be 0.0. Two assistants scored idea units in the recall responses from a subset of participants; their consistency in scoring was high (97%); hence, one assistant scored the remaining responses.

Students' correct identification of idea units in their prejudgment recall was relatively high ($M = 0.83$, $SE = 0.04$). Incorrect identifications were lower ($M = 0.30$, $SE = 0.03$), although significantly greater than 0.0, $t(30) = 9.6$. We also evaluated the degree to which identification errors were related to overconfidence by correlating incorrect identifications with self-score judgments for commission errors. The correlation was positive, $r = .37$, $p < .05$, suggesting that students' difficulties making accurate idea-unit judgments translated into less accurate self-score judgments. Thus, if middle school students could be trained to reduce their rate of incorrect identifications, overconfidence in their self-score judgments may be reduced even further.

Quality of control decisions based on self-score judgments. In Table 4, we report the proportion of definitions that would be dropped from study when students judged that they had correctly recalled the definition. For the full-definition and no-standard groups, the proportion of dropped definitions was based on students' self-score judgments indicating that a prejudgment recall response was fully correct (as in Experiment 1); for the idea-unit group, the proportion of dropped definitions was based on trials in which students indicated that all idea units were present in their prejudgment recall responses.

Most important, note that when decisions to drop definitions from study are based on students' idea-unit judgments, only 7% of the commission errors and 23% of the partially correct responses would be erroneously dropped from study. Even though this same criterion would mean that about 45% of the correctly recalled definitions would not be dropped, overlearning some already recalled definitions is arguably less problematic than failing to restudy definitions that have not yet been learned.

As in Experiment 1, we conducted a separate one-way ANOVA for each kind of recall response and found a significant effect of group for all three kinds of recall response, $F_s > 6.4$. Follow-up tests revealed that for commission errors, significantly fewer definitions would be dropped for the idea-unit group than for the full-definition group, $t(65) = 2.59$. Although the trend was in the expected direction, the proportion of commission errors that would be dropped did not differ between the full-definition and no-standard groups, $t(62) = 1.05$. For partially correct responses, fewer would be dropped for the idea-unit group than for the full-definition group, $t(58) = 4.55$; the comparison involving the full-definition and no-standard groups was not significant, $t(56) = 0.07$. Finally, for correctly recalled definitions, the idea-unit group would also drop significantly fewer than would either of the other groups, $t_s > 3.2$, and the difference between the full-definition and no-standard groups approached significance, $t(56) = 2.05$, $p = .045$.

General Discussion

The Promise of Standards for Improving Students' Judgment Accuracy

The major aim of this research was to assess whether standards of evaluation would reduce middle school students' overconfidence when judging the quality of their recall for previously studied key term definitions. We were particularly interested in their judgments for commission errors, where any self-score judgment other than "no credit" indicated overconfidence. In both experiments, full-definition standards reduced the overconfidence of middle school students' self-score judgments for commission errors. Moreover, the influence of the full-definition standards was selective: They reduced students' overconfidence in their commission errors and had a minimal influence on their self-score judgments of partially correct or correct responses. Thus, the standards did not influence students' scoring by merely making them more conservative—instead, the students better identified which recall responses were incorrect. The implication of these findings is straightforward. When students compare their answers with a full-definition standard, their judgments become better calibrated, which in turn should lead to more effective self-regulation of their learning (Winne, 2004). Fortunately, full-definition standards are readily available to students in the form of correct definitions that can be found in their textbooks, class notes, and other study materials.

To the extent that the benefits of full-definition standards for reducing overconfidence seems intuitive, the fact that students still show significant overconfidence when the correct answer is available is surprising. One possibility is that although full definitions do provide useful external standards, they do not provide sufficient feedback concerning what specific information is required for a response to be considered correct. Given this rationale, we hypothesized that idea-unit standards would further reduce overconfidence because they provide a more detailed standard of evaluation. In particular, idea-unit standards implicitly provide the smallest units of conceptual information that would warrant receiving any credit for a response. Unless students are forced to make such fine-grained distinctions, they may not even be aware that they are missing important components of a definition (cf. Goldsmith, Koriat, & Pansky, 2005). In support of this hypothesis, idea-unit standards significantly reduced the overconfidence of middle school students' judgments for their commission errors.

That this overconfidence could be further reduced was suggested by secondary analysis, which demonstrated that (a) students on average made incorrect identifications (i.e., they said an idea was present in their answer when it was not) on about 30% of the trials, and (b) across students, the proportion of incorrect identifications correlated positively with the degree of overconfidence for commission errors. One reason middle school students may make incorrect identifications is that they mark an idea unit as present if they believe they knew the correct answer even though it was not stated explicitly in their response. So, if a student recalls that a *biography* is "a book about something," perhaps the student would believe that *a book* implied that it is about a "true" story. Although we did warn students that they should mark an idea as present if and only if it was entirely and explicitly present in their response, the task instructions (which involved some practice with feedback) about how to make idea-unit judgments were brief. Perhaps more

extensive training on how to make accurate idea-unit judgments will help students reduce incorrect identifications.

Students' Evaluations as Input for Restudy Decisions

Accurate self-evaluation is important for developing techniques to support efficient and durable student learning. In the present case, if students judge that commission errors (or partially correct responses) are entirely correct, they may fail to restudy them. For instance, if they are working with an automated tutor that uses their self-evaluations to make decisions on scheduling further practice, the tutor would erroneously drop definitions from practice that a student actually has not learned well. In an automated tutor that we are currently developing, the tutor must use students' judgments to make restudy decisions on the basis of a student's current level of learning because reliable automated scoring methods are currently not available. Whereas computers can accurately score the quality of single-word responses and paragraph-length responses, current computer algorithms (e.g., based on latent semantic analysis) cannot accurately score the quality of sentence-length responses. Thus, improving students' accuracy for evaluating the quality of recalling definitions will be critical to the extent that poor accuracy will lead to dropping poorly learned definitions from study.

To estimate the extent of this problem, we computed the percentage of items that would be dropped from practice when students indicated that their recall responses were fully correct—that is, they made a self-score judgment of “full credit” (Experiments 1 and 2) or they judged that all the ideas in the correct definition were in their response (Experiment 2, idea-unit group only). Several outcomes were noteworthy (see Table 4). First, without any standard, more than 30% of commission errors would be dropped from study; such errors reflect the extreme overconfidence that students have in their commission errors when they cannot compare their answers with any standard. By contrast, full-definition standards reduced this rate of dropping commission errors, and with idea-unit standards, students (or an automated tutor) would drop only 7% of their commission errors. Furthermore, the idea-unit group would drop only 23% of their partially correct responses, which could also benefit from further study. Thus, if students want to minimize the likelihood of failing to study unlearned definitions, then they should use idea-unit standards to evaluate their learning.

Doing so must also be weighed against the fact that about 50% of the correctly recalled definitions would not be dropped from study (see Table 4, right column). Of the two possible errors (dropping unlearned definitions vs. restudying correctly recalled ones), we find the latter less worrisome from an educational perspective. First, dropping unlearned definitions from restudy is definitely not prudent; students would not have a chance to correct their errors, which in turn would lead to poor exam performance and a deficient knowledge base. Second, recent research has demonstrated that even correctly recalled items can benefit from further retrieval attempts (Karpicke & Roediger, 2007; Pyc & Rawson, 2007, 2009). That is, further practice with previously recalled definitions can increase long-term retention of those items. One intriguing possibility is that the optimal schedule of practice will change as a function of recall status; that is, definitions that are recalled correctly may benefit most from a long lag prior to the next study and retrieval attempt, whereas those that are partially

correct (or incorrect) may benefit most from shorter lags. If so, accurate student evaluations will be even more critical for making optimal decisions about how to schedule subsequent study and retrieval attempts.

Finally, if students allocate only a small amount of time to prepare for an exam, failing to drop already learned items from study may hamper overall performance because less time would be available to study unlearned items. Although possible, we note here that almost every cognitive technology aimed at improving student learning is time consuming and would be rendered largely ineffective when students decide to cram for an upcoming exam. That is, if students do not allocate enough time for studying overall, then no doubt many cognitive interventions recommended from educationally relevant research would be impotent. For these reasons, the potential benefits appear to outweigh the potential risks of using idea-unit standards to evaluate and guide learning.

Limitations and Future Directions

One potential limitation of the present research is that students may not have viewed the tasks as valuable, and hence they may have used minimal effort in learning the definitions and evaluating their recall. Although this minimal effort hypothesis is reasonable and could be evaluated in future research, we believe it is unlikely to account for the bulk of our findings. For instance, in both experiments, students knew that the materials they were studying would be relevant to course content, and hence they were aware of the value of the learning task. Moreover, even if the students did not use maximal effort in evaluating the recall responses, it is important to note that standards reduced overconfidence. Thus, students apparently were sufficiently motivated to demonstrate some benefits of standards as they evaluated the quality of their recall.

Another potential limitation is that our present research did not investigate the degree to which standards could help students to judge how well they *comprehend* each definition. Instead, we focused on students' judgments of the quality of their recall of definitions. Of course, students' ability to accurately judge their recall is important in its own right because doing so can help them learn to correctly recall the definitions, which in turn will be useful as they answer comprehension-based questions; that is, they must be able to remember a definition to apply it (e.g., to make an inference or to solve a problem). Even so, it will also be important to develop techniques that improve students' ability to evaluate their comprehension of classroom definitions (Dunlosky & Lipko, 2007; Wiley, Griffin, & Thiede, 2005). Given that idea-unit standards provide the smallest unit of conceptual information for each definition, they may hold promise for promoting students' evaluation of their comprehension. If so, such boosts in comprehension may also provide another means by which idea-unit standards improve students' evaluations of the quality of their recall. We leave evaluation of these possibilities for future research.

Finally, given the promise of idea-unit standards for improving the accuracy of students' evaluations, it is unfortunate that idea-unit standards are not usually available to students. Even so, the use of idea-unit standards could be readily adopted in the

classroom and in textbooks. Teachers could provide students with idea units for key concepts in a study guide or during a review session. Also, the end-of-the-chapter reviews in textbooks could provide answer keys including each key term definition along with the main ideas within it. Another possibility would be to teach students how to determine the idea units of key concepts so that they could develop their own idea units and use them to evaluate the quality of their recall. In a recent study, college students were able to generate idea units of key term definitions and successfully use them to accurately identify commission errors (Hartwig & Dunlosky, 2009). If middle school students are capable of developing this skill and it transfers across content domains, such training could provide a powerful metacognitive tool for helping them evaluate and regulate their learning. Regardless of how idea-unit standards are provided to students, using them shows promise for improving the efficacy of students' learning by helping them identify those concepts that they need to spend more time studying.

References

- Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*, 432–447.
- Baker, J., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review*, *13*, 60–65.
- Baker, L. (1984). Spontaneous versus instructed use of multiple standards for evaluating comprehension: Effects of age, reading proficiency, and type of standard. *Journal of Experimental Child Psychology*, *38*, 289–311.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Dunlosky, J., Hertzog, C., Kennedy, M. R. T., & Theide, K. W. (2005). The self-monitoring approach for effective learning. *Cognitive Technology*, *10*, 4–11.
- Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*, 228–232.
- Dunlosky, J., Rawson, K. A., & Hacker, D. (2002). Metacomprehension of science text: Investigating the levels-of-disruption hypothesis. In J. Otero & A. Graesser (Eds.), *The psychology of science text comprehension* (pp. 255–280). Hillsdale, NJ: Erlbaum.
- 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*, 551–564.
- Goldsmith, M., Koriat, A., & Pansky, A. (2005). Strategic regulation of grain size in memory reporting over time. *Journal of Memory and Language*, *52*, 505–525.
- Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, *77*, 334–372.
- Hacker, D. J. (1997). Comprehension monitoring of written discourse across early-to-middle adolescence. *Reading and Writing: An Interdisciplinary Journal*, *9*, 207–240.
- Hartwig, M. K., & Dunlosky, J. (2009, April). *Using idea-unit standards to improve students' self-monitoring of learning*. Paper presented at the Midwestern Psychological Association Conference, Chicago.
- Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach*. Orlando, FL: Harcourt Brace Jovanovich.
- Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162.
- Kikas, E. (1998). The impact of teaching on students' definitions and explanations of astronomical phenomena. *Learning and Instruction*, *8*, 439–454.
- 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. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370.
- Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology*, *103*, 152–166.
- 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.
- Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, *52*, 463–477.
- Miles, J. R., & Stine-Morrow, E. A. L. (2004). Adult age differences in self-regulated learning from reading sentences. *Psychology and Aging*, *19*, 626–636.
- Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 223–232.
- Nelson, T. O., Dunlosky, J., Graf, E. A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*, 207–213.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.
- 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.
- Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning*, *1*, 159–179.
- Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*, 1917–1927.
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437–447.
- Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 69–80.
- Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*, *19*, 559–579.
- Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, *28*, 1004–1010.
- Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: Evidence for developmental trends. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 391–409). New York: Taylor & Francis.
- Schneider, W., & Pressley, M. (1997). *Memory development between two and twenty* (2nd ed.). Mahwah, NJ: Erlbaum.
- Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204–221.
- Thiede, K. W. (1999). The importance of monitoring and self-regulation during multi-trial learning. *Psychonomic Bulletin & Review*, *6*, 662–667.
- Thomas, A. K., & McDaniel, M. A. (2007). Metacomprehension for educationally relevant materials: Dramatic effects of encoding-retrieval interactions. *Psychonomic Bulletin & Review*, *14*, 212–218.

- Wiley, J., Griffin, T., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology, 132*, 408–428.
- Winne, P. H. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research, 41*, 466–488.
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah, NJ: Erlbaum.
- Zabucky, K., & Moore, D. (1989). Children's ability to use three standards to evaluate their comprehension of text. *Reading Research Quarterly, 24*, 336–252.

Appendix

Terms, Definitions, and Idea Units

Term	Definition	Idea units (if used)
Genetics		
<i>Meiosis</i>	A process that requires two cell divisions, and the chromosomes only copy once in the formation of sex cells or gametes	(1) Process requiring two cell divisions (2) Chromosomes copy once (3) Formation of sex cells or gametes
<i>Homozygous</i>	The inheritance of two dominant genes or alleles for a specific trait	(1) Inheritance of two dominant genes (2) for a specific trait
Literary nonfiction		
<i>Primary source</i>	Any firsthand account recorded by people taking part in or witnessing an event	(1) A firsthand account (2) recorded by people who took part in the event (3) or who witnessed the event
<i>Biography</i>	A true story of a person's life written by someone else	(1) A true story (2) about a person (3) written by someone else

Received March 11, 2008

Revision received August 19, 2009

Accepted August 24, 2009 ■

New Editors Appointed, 2011–2016

The Publications and Communications Board of the American Psychological Association announces the appointment of 3 new editors for 6-year terms beginning in 2011. As of January 1, 2010, manuscripts should be directed as follows:

- *Developmental Psychology* (<http://www.apa.org/journals/dev>), **Jacquelynne S. Eccles, PhD**, Department of Psychology, University of Michigan, Ann Arbor, MI 48109
- *Journal of Consulting and Clinical Psychology* (<http://www.apa.org/journals/ccp>), **Arthur M. Nezu, PhD**, Department of Psychology, Drexel University, Philadelphia, PA 19102
- *Psychological Review* (<http://www.apa.org/journals/rev>), **John R. Anderson, PhD**, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213

Electronic manuscript submission: As of January 1, 2010, manuscripts should be submitted electronically to the new editors via the journal's Manuscript Submission Portal (see the website listed above with each journal title).

Manuscript submission patterns make the precise date of completion of the 2010 volumes uncertain. Current editors, Cynthia García Coll, PhD, Annette M. La Greca, PhD, and Keith Rayner, PhD, will receive and consider new manuscripts through December 31, 2009. Should 2010 volumes be completed before that date, manuscripts will be redirected to the new editors for consideration in 2011 volumes.