Title
Improving Metacognition in the Classroom

Journal
Zeitschrift für Psychologie, 229(2)

ISSN
2190-8370

Authors
Hausman, Hannah
Myers, Sarah J
Rhodes, Matthew G

Publication Date
2021-04-01

DOI
10.1027/2151-2604/a000440

Review Article

Improving Metacognition in the Classroom

Hannah Hausman[1], Sarah J. Myers[2], and Matthew G. Rhodes[2]

[1]Psychology Department, University of California, Santa Cruz, CA, USA
[2]Psychology Department, Colorado State University, Fort Collins, CO, USA

Abstract: Students are often overconfident in educational settings and struggle to differentiate between well-learned and poorly-learned concepts. The present article reviews current research on strategies that help students assess their understanding, with a focus on research using authentic educational tasks and materials. We propose a framework for these strategies that we refer to as *wait-generate-validate*. The wait-generate-validate strategies can give students a more objective measure of their learning from lectures, understanding of course concepts, text comprehension, problem-solving ability, and test preparedness. These strategies have been shown to lead to more effective study decisions and greater learning. Lastly, we translate the reviewed research into practical tips for students and teachers and conclude with recommendations for future research regarding how students judge their learning in diverse educational contexts.

Keywords: overconfidence, metacognition, metacomprehension, monitoring, judgments of learning

Most instructors have experienced a student attending office hours to discuss a disappointing test grade. It is frustrating for both the student and the teacher when a student works hard but earns a lower grade than expected. Metacognition refers to “one’s knowledge concerning one’s own cognitive processes and products or anything related to them,” or thinking about one’s thinking (Flavell, 1979). Metacognition consists of two key components: *monitoring* and *control*. Monitoring refers to the process of self-assessing one’s learning, understanding, or skill; control refers to self-regulated learning behaviors made as a result of monitoring (Dunlosky et al., 2016; Nelson, 1996; Nelson & Narens, 1990; Rhodes, 2019). When students misjudge how well they understand course content, it is an example of poor monitoring and may lead students to engage in poor control, such as not reviewing enough or reviewing the wrong material.

Research suggests that across a range of tasks and settings, students consistently overestimate their level of knowledge (e.g., Dunning et al., 2004) and struggle to differentiate among topics that they understand well and those they do not understand well (e.g., Dunlosky & Lipko, 2007; Thiede et al., 2009). Helping students improve their ability to judge their learning is important because it affects how students study and the amount they learn (e.g., Kornell & Finn, 2016).

This paper reviews teaching and study strategies that enhance students’ abilities to assess their understanding in educational settings. There is a large body of literature (e.g., Rhodes & Tauber, 2011) examining strategies to improve metacognitive monitoring accuracy in the laboratory, including with simple materials (e.g., word pairs such as *dog-table*). This review will focus on research conducted in classrooms or in laboratories, but with authentic educational tasks (e.g., watching video lectures, reading textbooks, learning foreign languages, etc.). Although we have included research conducted with students in kindergarten through college, discussing changes in metacognitive monitoring across the lifespan is outside the scope of this review (for a review of the development of metacognition in young children and adolescents, see Von der Linden et al., 2017). In addition to reviewing research on improving monitoring, we offer suggestions for how students and teachers can translate this research into practice.

Measuring Metacognitive Monitoring Accuracy

Metacognitive monitoring accuracy is measured by asking students to judge their level of learning; this judgment is then compared to their performance on a test (Rhodes, 2016; Schraw, 2009). Students’ judgments of their learning (JOLs) can be solicited in multiple ways. Students may be asked to make a *global JOL* by predicting the percentage of questions they will answer correctly on an upcoming test or rating their understanding of an entire text or concept on some scale. For example, a student may make the global JOL that she will earn a 95% on a unit exam or rate that

https://doi.org/10.1027/2151-2604/a000440

she understood her reading assignment a 3 out of 5. In addition to, or instead of providing a global JOL, students may be asked to make *item-by-item JOLs* by predicting the likelihood that they will remember a particular piece of information or correctly solve a particular type of problem on an upcoming test. For example, a student may study 20 anatomy terms and rate the likelihood on a scale from 0% to 100% that she will remember each of the 20 terms on an anatomy quiz.

There are two primary methods to determine the accuracy of students' metacognitive monitoring: *calibration* and *resolution* (e.g., Rhodes, 2016). Calibration refers to the absolute accuracy of metacognitive monitoring or the degree to which one's overall level of performance corresponds to one's predicted level of performance. Calibration can be calculated by taking the difference between actual test performance and a student's global JOL or by taking the difference between actual test performance and the average of a student's item-by-item JOLs. Consider a student who earns an 82% on an exam. If she predicted she would earn 80% or if her average item-by-item JOLs was an 80%, then the student is well-calibrated. She would be poorly calibrated and overconfident, if she predicted she would earn a 95% or underconfident if she predicted she would earn a 70%.

Resolution refers to the relative accuracy of metacognitive monitoring or the degree to which item-by-item JOLs differentiate between information that is remembered and information that is not remembered on a later test. Resolution is typically measured with the Kruskal-Goodman gamma correlation between item-by-item JOLs and item-by-item accuracy on a test, which can range from −1.0 to +1.0 (Nelson, 1984; but see Benjamin & Diaz, 2008; Masson & Rotello, 2009, for criticisms). A positive correlation indicates that information that was subsequently remembered tended to be given higher item-by-item JOLs than information that was not remembered. A more positive correlation reflects more accurate metacognitive monitoring. Gamma correlations less than .3, between .3 and .7, and greater than .7 can be considered low, moderate, and high resolution, respectively (Dunlosky & Lipko, 2007). A negative correlation indicates that information that was subsequently remembered tended to be given lower item-by-item JOLs than information that was not remembered. A more negative correlation reflects less accurate metacognitive monitoring. Near-zero correlations indicate little association between item-by-item JOLs and subsequent memory. Compare students A and B who both indicate that there is an 80% probability they will recall the term *epithelial* and only a 20% probability that they will recall the term *autonomic nervous system* on an upcoming anatomy quiz. Imagine student A remembers epithelial but not autonomic nervous system and student B remembers autonomic nervous system but not epithelial. Student A would show a more positive gamma correlation, and thus better resolution, compared to student B.

Metacognitive Monitoring With Educational Tasks

The Importance of Accurate Metacognitive Monitoring

Accurate metacognitive monitoring is important for students to effectively decide how much to study and what to study in order to maximize learning (for reviews, see Griffin et al., 2013; Koriat, 2012; Kornell & Finn, 2016; Metcalfe, 2009). When students have poor calibration and are overconfident in their understanding, they stop studying too early (Dunlosky & Rawson, 2012; Karpicke, 2009; Kornell & Bjork, 2008). For example, once students are able to correctly recall the definition on a flashcard, they overestimate their learning and assume they will continue to be able to recall the definition in the future. Consequently, students tend to drop flashcards from further studying after one successful recall attempt (Karpicke, 2009; Kornell & Bjork, 2008). However, students learn significantly more when they do not drop flashcards and continue practicing retrieving definitions beyond the first successful recall attempt (cf. Bahrick, 1979; Rawson et al., 2013).

When students have poor resolution, they cannot use their study time efficiently to prioritize reviewing information that is not as well learned (Little & McDaniel, 2015; Rawson et al., 2011; Thiede et al., 2003; Thomas & McDaniel, 2007a; Wiley et al., 2016). For example, Thiede and colleagues (2003) had participants read multiple text passages on different topics and rate their comprehension of each passage. When participants were supported in accurately identifying which passages they comprehended well and which passages they did not comprehend, participants learned more because they spent most of their review time rereading passages they had initially judged as poorly understood. In short, metacognitive monitoring significantly affects study decisions and thus learning.

Metacognitive Monitoring Accuracy: Typical Findings

Despite the importance of accurate metacognitive monitoring for study decisions and learning, research has consistently demonstrated that students are overconfident in educational settings. Students tend to be overconfident in how well they comprehend course content from reading assignments (Glenberg & Epstein, 1985, 1987; Thomas & McDaniel, 2007a; Wiley et al., 2016, 2018; Winne & Jamieson-Noel, 2002) and lectures (Falchikov & Boud, 1989; Szpunar et al., 2014). For example, Szpunar and colleagues (2014) found that students who watched an introductory statistics lecture video predicted that they would earn a 78% on a test of the content; they actually earned a 48%.

As a result of overconfidence while learning course content, students of all ages frequently predict that they will earn higher test scores and overall course grades than they actually do (Bol et al., 2005; De Bruin et al., 2017; Foster et al., 2017; Grimes, 2002; Hacker et al., 2000; Hartwig & Dunlosky, 2017; Karjanto & Yong, 2013; for reviews, see Hacker & Bol, 2019; Hacker, Bol, & Keener, 2008). This pattern of overconfidence does not necessarily diminish with experience, either. Foster and colleagues (2017) had students enrolled in a statistics course predict their performance before each exam in the course. Across all 13 exams, students were overconfident, predicting their scores would be an average of 7 percentage points higher than they actually were. Furthermore, students' metacognitive monitoring accuracy did not improve across the exams (see also Bol et al., 2005; Hacker, Bol, & Bahbahani, 2008; Hacker et al., 2000; Lipko, Dunlosky, & Merriman, 2009; Nietfeld et al., 2005, 2006; Saenz et al., 2019). Thus, teachers and students should actively engage in instructional and study strategies specifically designed to improve metacognitive monitoring; passively expecting monitoring to improve throughout a course is unlikely to help students assess their learning and make effective study decisions accordingly.

In addition to being overconfident in their overall level of understanding, students struggle to differentiate between well-learned and poorly-learned concepts. For example, when students read multiple texts, their judgments of comprehension often show low resolution (the average gamma correlation is approximately 0.30; Dunlosky & Lipko, 2007; Griffin et al., 2019; Maki, 1998; Thiede et al., 2009). Students' poor resolution is not limited to reading assignments, though, and can be seen at a much broader level in a course. Hartwig and Dunlosky (2017) gave students in an introductory statistics class a list of exam topics before the exam and had students predict the percentage of questions for each topic that they would answer correctly on the exam. Students' predictions showed low resolution: the gamma correlation between predicted and actual performance for different exam topics was 0.25 or lower.

In sum, it is essential for teachers to design classes and assignments that help students determine what they know and do not know. Furthermore, students should use strategies to accurately assess their learning as they read, listen to lectures, practice problems, and study for exams. This paper reviews instructional and study strategies that have been shown to improve metacognitive monitoring in educational settings. We also include practical suggestions for students and teachers based on prior research.

Improving Metacognitive Monitoring

How Students Make Metacognitive Judgments

The wealth of evidence that students have poor calibration and resolution demonstrates that students do not have direct access to the strength of their memories or the quality of their understanding. Instead, students must infer their level of knowledge based on cues available in the learning environment (Koriat, 1997), including the subjective experience of how easy material is to process or how familiar they are with the material as well as beliefs about how different instructional and study strategies affect learning (for reviews, see Bjork et al., 2013; Finn & Tauber, 2015).

One reason that metacognitive monitoring is often inaccurate is because the cues students use to judge their understanding of the material are unrelated to long-term learning and thus not related to future test performance. For example, students mistakenly believe they have learned more from a teacher with a very fluent presenting style compared to a teacher whose speech is more halted, even when the teachers present the same information (Carpenter et al., 2013, 2016). Similarly, making a reading slightly more difficult to process just by removing some letters from words can lower comprehension ratings, although this does not influence actual comprehension (Rawson & Dunlosky, 2002, Experiment 4).

However, it is not the case that students are incapable of assessing their own learning; students' metacognitive judgments are not fated to poor calibration and resolution. Instead, the accuracy of students' metacognitive judgments depends on the degree to which the cues used to infer learning are predictive of performance on a future test (Koriat, 1997). Consider a student who just attended a lecture and judges how well she understands the material to decide whether to attend a study group to prepare for an exam. The cues she could use to judge her learning under these conditions are likely not predictive of what she will remember or the questions she will be able to answer on the exam. For example, if the instructor is a smooth, fluent speaker, the student could mistake the ease with which she listened to the lecture with how well she

Table 1. A summary of the wait-generate-validate strategy

	Step 1: Wait	Step 2: Generate	Step 3: Validate
What it is	Wait minutes to days after learning material.	Then try to generate the material from memory.	Finally, validate the accuracy and completeness of the information you generated. Now you can judge your understanding of the material.
Why it works	Waiting will give you a more accurate sense of what information you can actually remember long enough for the test, rather than what information is just fresh on your mind.	Generating will give you a more accurate sense of what information you can produce on your own, without the help of your notes or book, just like you will have to do on a test.	Validating will give you a more accurate sense of how much you know, whether the information you can recall is correct, and whether you still have any misconceptions.
An example of how to do it	Wait until the day after a lecture to judge your understanding.	Then try to write down the key terms and concepts from the lecture without looking at your notes.	Finally, check your notes to verify that the terms and concepts you recalled were correct. Make sure you did not miss any major terms or concepts.

Note. For accurate metacognitive monitoring, students should use the wait-generate-validate strategy after learning material but before judging how well they understand it.

understands the concepts being taught (e.g., Carpenter et al., 2016). In addition, if she was familiar with some of the terms from a previous course, the student could mistake her familiarity with the terms with actually being able to define and recognize examples of the terms on the test (Glenberg & Epstein, 1987). Finally, the student might not attend the study group because she can easily recall facts or dates from the lecture, but this assessment will not be indicative of her test performance if she forgets the material over time or if the test involves conceptual questions requiring application and explanation (Thomas & McDaniel, 2007a, 2007b).

In contrast, if the student waits 2 days and then completes a practice exam from a previous semester without looking at her notes, the cues she could use to judge her understanding are much more likely to be predictive of what she will remember on the exam. These practice test conditions align more closely with the conditions of the actual test (Thomas & McDaniel, 2007a, 2007b), so her metacognitive monitoring judgments would better reflect what she actually retained from the lecture.

Thus, our empirically-based practical suggestions for students and teachers are based on an underlying principle: Students' metacognitive monitoring will be most accurate when they predict their future performance under conditions that closely align with how they will eventually be tested on that material. Specifically, we suggest that an activity will optimally support accurate metacognitive monitoring when it involves a 3-step process we refer to as *wait-generate-validate* (for similar strategies, see McDaniel et al., 2009; Rawson & Dunlosky, 2013; Rhodes et al., 2020). *Wait* refers to the suggestion that students should not judge their understanding immediately after learning, giving memories an opportunity to consolidate. After time passes, students should test their knowledge by actively *generating* information using tools such as flashcards or practice tests. Finally, in order to accurately assess their knowledge, students need to *validate* what they generated (i.e., check whether the information they generated was correct and complete). Table 1 summarizes the wait-generate-validate strategies. An additional benefit of the wait-generate-validate method is that these three strategies have also been shown to significantly improve learning (e.g., spacing: Cepeda et al., 2006; retrieval practice: Rowland, 2014; feedback: Hattie & Timperley, 2007).

Wait

Why Waiting Works

The first step of the wait-generate-validate method is to have students wait to judge their understanding after learning or reviewing information (for a meta-analysis, see Rhodes & Tauber, 2011). Just as an exam will not be administered immediately after information was covered in a class, students should not judge their learning immediately after studying. If, for example, students just learned the location of the four lobes of the brain, students may mistakenly believe that they will remember the lobes on the test in 2 weeks because the lobes are easy to label at that moment. However, how easy it *feels* to recall information or solve a problem in the moment is not indicative of long-term learning (cf. Benjamin et al., 1998). Judgments of learning made immediately after learning or studying are often not well-aligned with future performance because the content is fresh in one's mind such that what is accessible at the moment is not predictive of what will be recallable later after forgetting from long-term memory happens. However, when judgments are delayed, the ease or difficulty with which information can be remembered is much more diagnostic of long-term learning, thereby supporting more accurate predictions of future performance.

How to Wait

Waiting to judge learning has been demonstrated to improve metacognitive monitoring accuracy – particularly resolution – across a range of tasks, including foreign languages (Scheck et al., 2004; Thiede & Dunlosky, 1994), the meaning of common idioms (Van Loon et al., 2013a), and text passages (Anderson & Thiede, 2008; Thiede & Anderson, 2003; Thiede et al., 2003, 2005; but see Maki, 1998). For example, Thiede and Dunlosky (1994) had participants learn Swahili-English translations. Either immediately after studying the translation or several minutes later, participants were presented with the Swahili word and were asked to rate the likelihood they could recall the English translation on a later test. Delaying judgments by even a few minutes dramatically improved the resolution of item-by-item judgments to the point where students could almost perfectly differentiate between well-learned and poorly-learned translations (gamma > 0.8).

Although research suggests that students should wait after learning or studying to accurately judge their level of knowledge, there has not been a systematic investigation of how long they should wait (for related research on short time scales, see Kelemen & Weaver, 1997). For practical purposes, we recommend that students wait approximately 1 day to judge their learning. If students need to quickly identify which information to continue studying, though, waiting even a few minutes to judge one's learning can improve metacognitive monitoring accuracy (e.g., Thiede & Dunlosky, 1994).

Getting the Most Out of Waiting

One note of caution: Although waiting to judge learning is a useful tool for improving monitoring accuracy, it is not effective under all circumstances. Waiting improves monitoring accuracy only to the degree that it yields cues that are more diagnostic of future memory performance than the cues that come from making immediate judgments. Waiting is beneficial because when students wait to judge their learning and they query their memory, the information that can be recalled then is typically a strong predictor of what will be recalled on subsequent tests (Rhodes & Tauber, 2011). Therefore, the key to waiting is that students must actually query memory to check their current knowledge (Dunlosky & Nelson, 1992; Thiede et al., 2005, Experiment 4; Van Loon et al., 2013a). For example, Thiede and colleagues (2005) had participants read a series of texts on different topics (e.g., sleep, the Titanic). After the delay of reading all of the texts, participants either read keywords or generated their own keywords for each text (e.g., iceberg, shipwreck, tragedy) and then rated their comprehension of the text on a scale from 1 to 7. Finally, participants took a test on all of the texts. The comprehension ratings, which were made after a delay, showed significantly better resolution when participants had to generate their own keywords rather than reading provided keywords. Thus, waiting seems to be most effective for assessing one's learning when it is combined with a prompt to encourage students to generate information, as described in more detail in the next section.

Generate

Why Generation Works

The value of generation for metacognitive monitoring is that it allows students to evaluate their level of knowledge based on how much they can generate or recall at the moment. The experience of failing to generate information can lead learners to lower their predictions of overall future performance, thereby tempering overconfidence (e.g., Baars, Van Gog, et al., 2014; Baars et al., 2017; Miller & Geraci, 2014, 2016; Thomas et al., 2016). Furthermore, generation activities also improve resolution, helping students identify the content that is well-learned and poorly-learned based on generation success (Ariel & Dunlosky, 2011; Finn & Metcalfe, 2007, 2008). For example, Jacoby and colleagues (2010) had participants learn to classify exemplars of different families of birds (e.g., Orioles, Jays, Thrashers, etc.) and then predict how well they could classify new exemplars from each family. Participants' performance predictions for each family were almost perfectly correlated with their accuracy in generating family names for exemplars during practice. This suggests that students' predictions are strongly based on the success of their generation or retrieval attempts.

How to Use Generation

There is a range of activities that students can engage in to improve metacognitive monitoring because they involve attempting to recall information. Flashcards are a simple and common method for students to practice generating definitions, translations, or facts (Wissman et al., 2012), which can significantly improve metacognitive monitoring accuracy. For example, Kornell and Rhodes (2013) had participants learn English-Indonesian translations. After studying all of the translations, participants either restudied the English-Indonesian pair or were presented only the English term and attempted to recall the Indonesian word. After restudying or attempting to recall the translation, participants rated the likelihood they could recall the Indonesian term on a later test. Participants' judgments showed significantly better resolution following recall attempts than restudying, and participants made better study decisions accordingly. Compared to participants who could practice recalling the Indonesian word, participants who restudied the translations opted to drop many terms from restudying that they did not actually know on the final test.

Practicing recall is also an effective tool for monitoring learning of more complex content than definitions or foreign-language translations, such as how to solve problems in math and the sciences (but see Bol & Hacker, 2001). When students study worked-out examples of how to solve math or science problems, they can be overconfident in their ability to independently solve similar problems in the future. However, overconfidence can be reduced and students can become much better calibrated if they have to try to solve example problems on their own before predicting their future test performance (Baars, Van Gog, et al., 2014; Baars et al., 2017). In some cases, though, solving practice problems can reduce overconfidence so much it produces underconfidence (Baars et al., 2013).

Although a growing body of research is examining the role of generation in metacognitive monitoring of problem-solving ability, the majority of research on improving metacognition with more complex materials has focused on improving judgments of reading comprehension, or metacomprehension (for reviews, see Dunlosky & Lipko, 2007; Dunlosky et al., 2002; Griffin et al., 2019; Thiede et al., 2009; Thomas & McDaniel, 2007a). Students can improve the resolution of their metacomprehension by completing a reading and then recalling definitions (Dunlosky et al., 2005), recalling key ideas (Little & McDaniel, 2015), generating keywords (Thiede et al., 2003, 2012), summarizing (Thiede & Anderson, 2003; Thiede et al., 2010) or explaining (Fukaya, 2013) the text from memory, creating concept maps (Redford et al., 2012; Thiede et al., 2010; Van Loon et al., 2014), or returning to the text and explaining what each paragraph adds to the meaning of the text as a whole (Griffin et al., 2008).

Similar to summarizing, students can also more accurately assess their knowledge by explaining concepts in detail from memory. Rozenblit and Keil (2002) demonstrated that participants were overconfident in how well they understood complex processes or devices such as how ocean tides work or how car batteries store electricity. However, participants' judgments of their understanding were significantly better calibrated once they were asked to explain the process or device in depth; participants reported being surprised by how shallow their understanding really was. One explanation for participants' initial overconfidence is that they based their confidence on the visible components that they were familiar with (e.g., a car battery is a box with a positive and negative side and contains fluid) and were not aware of all the hidden mechanistic features (e.g., the fluid is an electrolyte that allows for electrons to be transferred from the anode chemical to the cathode chemical through an oxidation-reduction reaction; Rozenblit & Keil, 2002). In other words, participants' overconfidence may have stemmed from judging their understanding at too abstract of a level and did not focus enough on the details, unless prompted to (Alter et al., 2010). This explanation has important implications for the classroom. Many natural phenomena that students learn in school (e.g., photosynthesis, digestion, etc.) have visible components (e.g., leaves, roots, the sun), but are explained by hidden causal mechanisms (e.g., ATP, carbon dioxide, light energy), making it likely that students will overestimate their understanding of such complex processes unless prompted to generate detailed explanations.

Getting the Most Out of Generation

In order for generation activities to support accurate metacognitive monitoring, performance on the generation activity should be predictive of performance on the test students are preparing for. At least two factors are important. First, students should wait after learning to generate because what can be recalled after a delay is typically more diagnostic (predictive) of long-term memory than what can be recalled immediately after learning (Anderson & Thiede, 2008; Thiede & Anderson, 2003; Thiede et al., 2003; Van Loon et al., 2013a, 2014). For example, Thiede and colleagues (2003) had participants read texts on a range of topics (e.g., the effects of alcohol on sleep, Norse settlements, the Titanic, etc.) and then generate keywords (e.g., iceberg, shipwreck, tragedy, etc.) for each text from memory. Participants generated keywords immediately after reading each text, waited to generate keywords until they had read all of the texts, or did not generate any keywords. Participants then rated their comprehension of the texts and took an initial test. On the initial test, participants performed similarly regardless of whether or not they generated keywords. However, participants who waited to generate keywords were superior at predicting their initial test performance. Specifically, the comprehension judgments showed significantly better resolution following the delayed keywords compared to immediate keywords; generating keywords immediately after reading did not improve monitoring accuracy in comparison to not generating keywords.

After the initial test, Thiede and colleagues (2003) gave participants the opportunity to reread the texts before a final test. As a result of being able to better discriminate between well-understood and poorly-understood texts, participants who generated keywords after a delay learned more from the review because they prioritized rereading the texts they understood the least. On the final test, participants who generated delayed keywords performed significantly better than participants who either generated keywords immediately or did not generate keywords. Thus, having students wait to generate information and then judge their learning has not only been shown to help students better identify what they know and do not know, but it has also been shown to help students make more effective subsequent study decisions as a result (De Bruin et al., 2011; Little & McDaniel, 2015; Thiede et al., 2003, 2012; Van Loon et al., 2013a; but see Kimball et al., 2012).

In addition to the timing of the generation activity, the format of the generation activity can influence its efficacy as a tool for supporting metacognitive monitoring. Importantly, the generation activity and the final test should require similar types of processing (Dutke et al., 2010; Thiede et al., 2012; Thomas & McDaniel, 2007a, 2007b) and cover the same content as the test students will receive. Note that the generation activity and the test do not need to be superficially similar to support accurate metacognitive monitoring (Dunlosky & Nelson, 1997; Thomas & McDaniel, 2007a). For example, if students are preparing for a multiple-choice test, they do not need to take a multiple-choice practice test to accurately judge their learning. Instead, if students will have to recognize definitions and examples of key concepts on a multiple-choice exam, they could judge their knowledge by practicing recalling definitions and generating examples of these concepts from memory using flashcards. In contrast, if students will have to write an essay comparing and contrasting psychological theories, they could judge their knowledge by practicing recalling the key ideas of each theory and explaining the similarities and differences to a classmate.

Therefore, it is important for students to find out what type of content and questions will be on a test in order to engage in generation activities that match the type of processing involved on the test and accurately assess learning (Griffin et al., 2019; Thiede et al., 2011; Wiley et al., 2016). Even just knowing whether they will be tested on memory for facts or deeper conceptual understanding can help students more accurately predict their test performance (Griffin et al., 2019; Thiede et al., 2011). In sum, to maximize metacognitive monitoring accuracy, students should wait and then judge their learning after engaging in generation activities that align with the final test students are preparing for.

Validate

Why Validating Works

After students have generated information, they should validate that the answers they generated are correct and complete. Making mistakes when generating information helps students identify what they have not learned well yet and need to continue reviewing. However, students are not always aware of when they have made mistakes (Dunlosky & Rawson, 2012; Händel et al., 2020; Lipko, Dunlosky, Hartwig, et al., 2009; Rawson & Dunlosky, 2007; Rawson et al., 2011; Van Loon et al., 2013b; Zamary et al., 2016). For example, Zamary and colleagues (2016) had students generate examples of previously learned psychology concepts. Students then judged their examples as incorrect, partially correct, or fully correct. Students showed only modest resolution (i.e., only a modest ability to differentiate among incorrect, partially correct, and fully correct examples). Furthermore, students were overconfident, being more likely to give full credit than no credit to answers that were entirely incorrect.

When students are overconfident in the quality of their answers and cannot differentiate between which answers were correct and incorrect, their subsequent studying is less effective, and long-term learning suffers: Students terminate studying prematurely and do not prioritize reviewing the information they understand the least well, which undermines long-term learning (Dunlosky & Rawson, 2012; Rawson et al., 2011; Van Loon et al., 2013b; but see Kornell & Rhodes, 2013).

How to Use Validation

Providing corrective feedback after students generate answers can help them assess the quality of their answers and thus evaluate their current knowledge (Baars, Vink, et al., 2014; Baker et al., 2010; Dunlosky & Rawson, 2012; Lipko, Dunlosky, Hartwig, et al., 2009; Rawson & Dunlosky, 2007; Rawson et al., 2011; but see Zamary et al., 2016). For example, most textbooks include a list of key terms at the end of each chapter. Practicing recalling the definitions of these terms, especially a day or more after reading the chapter, should help students accurately monitor what they learned from the chapter because it involves generation after a delay. However, prior research suggests that students will often be overconfident in the quality of the definitions they recall. That is, students may generate a definition that is only partially correct or completely incorrect, but mistakenly believes they correctly defined the term. To help identify which definitions were accurately recalled, students should carefully compare the definitions they generated with the definitions provided in the text. Similarly, when solving problems in math or the sciences, students should carefully compare their work to the solutions provided by the instructor, if solutions are available (Baars, Vink, et al., 2014).

Getting the Most Out of Validation

Although providing the correct answer can help students assess the quality of their own answers, students tend to remain overconfident (Dunlosky & Rawson, 2012; Rawson & Dunlosky, 2007; Lipko, Dunlosky, Hartwig, et al., 2009). Rawson and Dunlosky (2007) had students read texts and recall definitions for key terms from the texts. Even when students explicitly compared the definitions they generated to the correct definitions, students still thought many of their partially correct or incorrect answers were fully correct. One strategy that can be used to further improve the students' accuracy in judging the quality of their answers is to provide *idea-unit feedback* (Dunlosky & Rawson, 2012; Rawson et al., 2011; but see Zamary et al., 2016). Idea-unit feedback entails breaking down the correct answer into its key features and having students score whether their answer included each of the features or idea units. The idea units for the definition for the concept of *proactive interference* could be (1) newer information cannot be recalled because (2) similar (3) older information is getting in the way.

Critically, when students can more accurately judge the quality of their answers because of idea-unit feedback, their subsequent study choices become more strategic, which leads to larger improvements in long-term learning (Dunlosky & Rawson, 2012; Rawson et al., 2011). Rawson and colleagues (2011) had students read a series of medical information texts about how to prevent problems related to diabetes. Two days later, students judged their understanding by answering a series of practice test questions (e.g., "How do you prevent eye problems caused by diabetes?"). Students then judged the quality of their answers with or without idea-unit feedback and then chose which texts to reread. Students who judged the quality of their original answers alongside idea units were less overconfident, better calibrated, and showed greater resolution in differentiating among correct, partially correct, and incorrect answers. Interestingly, the group who judged their answers with idea-unit feedback actually reread fewer texts and for a shorter period of time yet learned more from rereading than the group who did not have idea-unit feedback. On a final test on the texts five days later, the idea-unit group outperformed the no idea-unit group by approximately 20 percentage points. The idea-unit feedback allowed students to better judge the quality of their practice test answers and thus identify which concepts they understood the least well so that they could focus on the relevant texts during rereading. Thus, providing feedback – particularly when it is broken down into idea units – helps students assess their understanding and make more effective subsequent study decisions.

Examples for Implementing Wait-Generate-Validate

The following suggestions are examples of how students and teachers can incorporate the wait-generate-validate strategies into various educational activities to help students accurately monitor their learning. The suggestions are not listed by priority, but rather, can be adopted and adapted by students and teachers to fit their educational needs and contexts. Furthermore, although they were written with students and instructors in higher education in mind they could easily be adapted for almost any grade level.

Suggestions for Students

Actively Generate Information; Do Not Passively Take It In

During the lecture, when the instructor or another student asks a question during lecture, generate your own answer to the questions, even if you do not plan on volunteering your answer to the class. As you read, stop to explain the information to yourself every few paragraphs. Think, "what were the key points from the last few paragraphs and what did each paragraph add to the meaning of the overall text?" If you are studying for an exam in a group, try recalling key information for yourself. Within the group, elaborate on each other's answers: explain how ideas are related or where another group member made a mistake.

Create Your Own Study Guides

If the lecture slides are available, create a shell of the lecture by taking out definitions, explanations, and examples, leaving only slide headers, key terms, and the outline of important diagrams. Later, test your understanding by working through the lecture shell as much as you can. Similarly, as you read, write down the headings to key sections and the names of key terms. Wait and then use these as cues to generate the information presented in the text.

Make Use of "Downtime" to Check What You Remember

When you are waiting for the bus, brushing your teeth, walking between classes, and so forth, try to remember information from your reading assignments and classes from the previous days. Generate keywords, summarize, explain an important process, draw diagrams or concept maps if you have paper, and so forth. Start this habit as soon as new information has been introduced; do not wait until the day before the test.

Structure Your Exam Review to Match the Test

If you have access to any practice questions or problems from lectures, the textbook, sample exams, or study guides, there is a good chance these questions and problems will align with the test you are preparing to take. Try answering these questions from memory and resist the urge to look at your notes. Once you have generated your own answer to every question, make sure your answers are correct and complete: use your notes and textbook, go to office hours, or compare answers with other students. If a study guide is not provided or is sparse, consider the types of questions that will be on the test. For example, will you be tested on your memory for details (e.g., key dates, names, locations)? Will you be tested on your ability to make inferences based on what you learned in class (e.g., predict what will happen in a chemical reaction), to recall definitions, generate examples, solve problems, or compare and contrast theories? Practice the same kinds of thinking as you review. You could even try writing your own exam questions in the format of the upcoming test.

Get It Right Multiple Times on Multiple Days

Recalling information once does not guarantee you will remember it on a test in the future. Engage in at least one of these generation activities on multiple days leading up to an exam. Be sure to correctly recall each piece of information approximately three times during *each* study session. For example, if you are practicing flashcards for a vocabulary quiz on Wednesday, aim to correctly recall each flashcard term correctly three times on Monday and Tuesday. Similarly, do not only use the study guide or practice problems a single time. Aim to correctly complete any study guide questions or practice problems on multiple days.

Suggestions for Teachers

Have Students Repeatedly Generate Content

Have students repeatedly revisit both recent content and content from earlier in the course via generation. For example, start each class with a low-stakes quiz that primarily covers content from the previous week, but also incorporates some material from the previous unit. Similarly, you could start each class by having students summarize the previous day's material to a classmate and identify how at least one concept from earlier in the course relates to that material. For a review activity, you could have students generate the key concepts from the unit from memory, draw a concept map, and then explain to a classmate how the concepts are related. As a rule of thumb, consider having students do some sort of generation activity with each key concept at least three times before an exam.

Make Generation the Default

If you pose questions in class, present a practice problem, or offer a study guide, students may wait until you offer the answer and avoid generating their own answer first. Therefore, encourage all students to write their answers down before you call on someone. You could also have students turn to a neighbor and each explains their thinking first. For study guides, do not put answers directly on any study guide. Withholding answers will encourage students to try to answer the questions from memory before looking at the answers.

Give Feedback

If students are going to generate information, be sure to give feedback and give students time to review the feedback. If you start class with a low-stakes quiz on the material from the previous class, go over the answer to each question before continuing with that day's lecture. If you incorporate an in-class activity in which students have to generate examples or explain their thinking, make sure that students' examples and explanations are complete and accurate. If you assign homework, offer detailed feedback or a solution key to the homework. Similarly, ensure that detailed answers to study guides are available (online, in a recitation, in-office hours, etc.).

Tell Students What They Need to Know

Students need not be told exactly which questions will be on an exam. However, it will benefit learning and metacognition to explicitly tell students what kinds of thinking they will be required to do on the test (e.g., defining terms, recognizing examples, comparing and contrasting theories, explaining processes, making predictions, etc.). For example, go beyond listing on the study guide the key concepts that will be on a test. Instead, tell students how they will be asked to use those concepts on the test, and better yet, give example questions or problems.

Make Activities Align With the Test

You do not have to ask the same question on a quiz and the exam. However, make sure the type of thinking required on quizzes, in-class activities, homework assignments, and study guides align with how students will be tested. If students will have to compare and contrast two theories on an exam, do not just quiz them on definitions in class. Instead, students could have a debate in class over which theory they believe is more appropriate. Similarly, if the exam will require students to differentiate between different types of math problems and select the appropriate formula, do not have each homework assignment be dedicated to a single type of problem, where students cannot practice identifying which formula they should use.

Future Directions

In addition to refining the wait, generate, and validate strategies, future research should examine other means by which to support students' metacognitive monitoring. Recent findings suggest that it would be particularly beneficial to investigate how the design of course materials and digital learning environments influence metacognitive monitoring accuracy.

Course Material Design

This review has covered strategies that teachers and students can implement for students to more accurately assess their knowledge. The strategies discussed suggest how students should interact with the course materials but say nothing about the design of the materials themselves. For example, students can better judge their understanding of reading assignments by waiting and generating keywords or summaries. However, it may also benefit student metacognition if teachers and textbook publishers alter course materials to provide students more diagnostic cues about their level of understanding, for example, by redesigning reading assignments to incorporate only informative diagrams and not decorative images, as most textbooks do.

Individual studies have examined course material designs that can interfere with accurate metacognitive monitoring, including adding decorative pictures (e.g., Cuevas et al., 2002; Jaeger & Wiley, 2014; Serra & Dunlosky, 2010) or analogies to texts (Wiley et al., 2018). Other studies have identified features of course materials that can support accurate metacognitive monitoring, including inserting questions into the text (Mitsuda, 1988; Walczyk & Hall, 1989) and adding sentences to texts that state and then refute common misconceptions (Prinz et al., 2018). However, across these different strands of research, no clear overarching principles have emerged for improving metacognitive monitoring. Future research should examine principles for course material design that support accurate monitoring and identify which principles are likely to have the biggest impact on metacognitive monitoring accuracy.

Digital Learning Environments

Given that instruction – including lectures, activities, homework, and entire courses – is increasingly moving online (e.g., Brown & Green, 2016), future research should focus on identifying the features of digital learning environments that optimize students' ability to monitor their learning and make effective follow-up study decisions (Azevedo & Aleven, 2013). Improving metacognition in digital learning environments is especially important given that research has revealed that monitoring accuracy tends to be lower using digital rather than paper materials (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Halamish & Elbaz, 2020; Sidi et al., 2017).

The scope of digital learning environments can range from individual assignments to full courses and includes multimedia reading assignments with links to glossaries, videos, and examples (e.g., Azevedo, 2005a); online math and science problems that offer video explanations of sample problems, hints, and step-by-step solutions (e.g., MyMathLab); online science experiment simulations (e.g., Rutten et al., 2012); intelligent tutoring systems, which adapt instruction in the moment to meet students' needs (e.g., Graesser, Conley, & Olney, 2012); and fully digital courses without a human instructor (e.g., ALEKS). When students cannot accurately monitor their learning in digital learning environments, they do not effectively make use of the available instructional resources, and learn less as a result (Mudrick et al., 2018).

Such digital learning environments offer access to abundant resources and typically give the student significant control over which resources to access and when. One decision that online learning platforms have to make is when and how to offer hints and solutions (e.g., Aleven et al., 2006; Azevedo, 2005b; Azevedo & Hadwin, 2005). Research suggests that students do not seek help optimally in digital learning environments (e.g., Aleven et al., 2006). Aleven and colleagues (2006) found that geometry students rapidly clicked through the hints to get to the solution on about one-third of problems; misusing help in ways like this was negatively correlated with learning.

Based on the value of generation, students may be able to more accurately assess their current understanding if they have to attempt to answer a question or solve a problem before viewing a hint or answer. Research has begun to identify promising ways to train students to make more effective use of help features in digital learning environments (e.g., Aleven et al., 2016; Azevedo & Aleven, 2013; Azevedo & Cromley, 2004; Azevedo & Hadwin, 2005; Feyzi-Behnagh & Azevedo, 2012; Feyzi-Behnagh et al., 2014; Roll et al., 2006). More broadly, future research should examine how much control to give students over access to hints and other resources; students should have the information they need to learn when they need it but should also have to engage in a generation to accurately assess their learning and make optimal study decisions.

Concluding Remarks

Students have the opportunity to take control of their learning outside of class. However, in order to make effective decisions about what to study and how much to study, students must be able to accurately assess their current level of understanding and differentiate between topics that are well-learned and poorly-learned. There are also myriad strategies within the wait-generate-validate framework that students can implement to more accurately judge their learning. At their core, the wait-generate-validate strategies emphasize that students should assess their learning under conditions that closely match how they will have to use their knowledge on exams in the future. Teachers can further support student metacognition by providing students with information about how they will be tested, what they will be tested on, and opportunities to practice. Finally, we encourage students and teachers to implement wait-generate-validate strategies because not only do they enhance metacognitive monitoring, but because they are also beneficial for long-term learning.

References

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied, 17*(1), 18–32. https://doi.org/10.1037/a0022086

Ackerman, R., & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior, 28*(5), 1816–1828. https://doi.org/10.1016/j.chb.2012.04.023

Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education, 16*(2), 101–128.

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education, 26*(1), 205–223. https://doi.org/10.1007/s40593-015-0089-1

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99*(3), 436–451. https://doi.org/10.1037/a0020218

Anderson, M. C., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica, 128*(1), 110–118. https://doi.org/10.1016/j.actpsy.2007.10.006

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39*(1), 171–184. https://doi.org/10.3758/s13421-010-0002-y

Azevedo, R. (2005a). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*(4), 199–209. https://doi.org/10.1207/s15326985ep4004_2

Azevedo, R. (2005b). Computer environments as metacognitive tools for enhancing learning. *Educational Psychologist, 40*(4), 193–197. https://doi.org/10.1207/s15326985ep4004_1

Azevedo, R., & Aleven, V. (2013). *International handbook of metacognition and learning technologies* (Vol. 26). Springer. https://doi.org/10.1007/978-1-4419-5546-3

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*(3), 523–535. https://doi.org/10.1037/0022-0663.96.3.523

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition – Implications for the design of computer-based scaffolds. *Instructional Science, 33*(5/6), 367–379. https://doi.org/10.1007/s11251-005-1272-9

Baars, M., Van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*(3), 382–391. https://doi.org/10.1002/acp.3008

Baars, M., Van Gog, T., de Bruin, A., & Paas, F. (2017). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology, 37*(7), 810–834. https://doi.org/10.1080/01443410.2016.1150419

Baars, M., Vink, S., Van Gog, T., De Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92–107. https://doi.org/10.1016/j.learninstruc.2014.04.004

Baars, M., Visser, S., Van Gog, T., De Bruin, A., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology, 38*(4), 395–406. https://doi.org/10.1016/j.cedpsych.2013.09.001

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*(3), 296–308. https://doi.org/ 10.1037/0096-3445.108.3.296

Baker, J. M., Dunlosky, J., & Hertzog, C. (2010). How accurately can older adults evaluate the quality of their text recall? The effect of providing standards on judgment accuracy. *Applied Cognitive Psychology, 24*(1), 134–147. https://doi.org/10.1002/acp.1553

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 73–94). Psychology Press. https://doi.org/10.4324/9780203805503

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55–68. https://doi.org/10.1037/0096-3445.127.1.55

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69*(2), 133–151. https://doi.org/10.1080/00220970109600653

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269–290. https://doi.org/10.3200/JEXE.73.4.269-290

Brown, A., & Green, T. (2016). Virtual reality: Low-cost tools and resources for the classroom. *TechTrends, 60*(5), 517–519. https://doi.org/10.1007/s11528-016-0102-z

Carpenter, S. K., Mickes, L., Rahman, S., & Fernandez, C. (2016). The effect of instructor fluency on students' perceptions of instructors, confidence in learning, and actual learning. *Journal of Experimental Psychology: Applied, 22*(2), 161–172. https://doi.org/10.1037/xap0000077

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review, 20*(6), 1350–1356. https://doi.org/10.3758/s13423-013-0442-z

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Cuevas, H. M., Fiore, S. M., & Oser, R. L. (2002). Scaffolding cognitive and metacognitive processes in low verbal ability learners: Use of diagrams in computer-based training environments. *Instructional Science, 30*(6), 433–464. https://doi.org/10.1023/A:1020516301541 de Bruin, A. B., Kok, E. M., Lobbestael, J., & de Grip, A. (2017). The impact of an online tool for monitoring and regulating learning at university: Overconfidence, learning strategy, and personality. *Metacognition and Learning, 12*(1), 21–43. https://doi.org/10.1007/s11409-016-9159-5

de Bruin, A. B., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*(3), 294–310. https://doi.org/ 10.1016/j.jecp.2011.02.005

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 23–37). Oxford University Press.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20*(4), 374–380. https://doi.org/10.3758/BF03210921

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language, 36*(1), 34–49. https://doi.org/10.1006/jmla.1996.2476

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 68–92). Cambridge University Press.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*(4), 551–565. https://doi.org/10.1016/j.jml.2005.01.011

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Dutke, S., Barenberg, J., & Leopold, C. (2010). Learning from text: Knowing the test format enhanced metacognitive monitoring. *Metacognition and Learning, 5*(2), 195–206. https://doi.org/10.1007/s11409-010-9057-1

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430. https://doi.org/10.3102/00346543059004395

Feyzi-Behnagh, R., & Azevedo, R. (2012). The effectiveness of a pedagogical agent's immediate feedback on learners' metacognitive judgments during learning with MetaTutor. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *International Conference on Intelligent Tutoring Systems* (pp. 651–652). Springer. https://doi.org/10.1007/978-3-642-30950-2_105

Feyzi-Behnagh, R., Azevedo, R., Legowski, E., Reitmeyer, K., Tseytlin, E., & Crowley, R. S. (2014). Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system. *Instructional Science, 42*(2), 159–181. https://doi.org/10.1007/s11251-013-9275-4

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(1), 238–244. https://doi.org/10.1037/0278-7393.33.1.238

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*(1), 19–34. https://doi.org/10.1016/j.jml.2007.03.006

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review, 27*(4), 567–586. https://doi.org/10.1007/s10648-015-9313-7

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning, 12*(1), 1–19. https://doi.org/10.1007/s11409-016-9158-6

Fukaya, T. (2013). Explanation generation, not explanation expectancy, improves metacomprehension accuracy. *Metacognition and Learning, 8*(1), 1–18. https://doi.org/10.1007/s11409-012-9093-0

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 702–718. https://doi.org/10.1037/0278-7393.11.1-4.702

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*, 84–93. https://doi.org/10.3758/BF03197714

Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA handbooks in psychology®. APA educational psychology handbook, Vol. 3. Application to learning and teaching* (pp. 451–473). American Psychological Association. https://doi.org/10.1037/13275-018

Griffin, T., Mielicki, M., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge handbook of cognition and education* (Cambridge Handbooks in Psychology, pp. 619–646). Cambridge University Press. https://doi.org/10.1017/9781108235631.025

Griffin, T. D., Wiley, J., & Salas, C. R. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 19–34). Springer. https://doi.org/10.1007/978-1-4419-5546-3_2

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*(1), 93–103. https://doi.org/10.3758/MC.36.1.93

Grimes, P. W. (2002). The overconfident principles of economics student: An examination of a metacognitive skill. *The Journal of Economic Education, 33*(1), 15–30. https://doi.org/10.1080/00220480209596121

Hacker, D., & Bol, L. (2019). Calibration and self-regulated learning: Making the connections. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge handbook of cognition and education* (Cambridge Handbooks in Psychology, pp. 647–677). Cambridge University Press. https://doi.org/10.1017/9781108235631.026

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*(2), 101–121. https://doi.org/10.1007/s11409-008-9021-5

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–455). Psychological Press. https://doi.org/10.4324/9780203805503

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

Halamish, V., & Elbaz, E. (2020). Children's reading comprehension and metacomprehension on screen versus on paper. *Computers & Education, 145*, Article 103737. https://doi.org/10.1016/j.compedu.2019.103737

Händel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing. *Learning and Instruction, 65*, Article 101245. https://doi.org/10.1016/j.learninstruc.2019.101245

Hartwig, M. K., & Dunlosky, J. (2017). Category learning judgments in the classroom: Can students judge how well they know course topics? *Contemporary Educational Psychology, 49*, 80–90. https://doi.org/10.1016/j.cedpsych.2016.12.002

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1441–1451. https://doi.org/10.1037/a0020636

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction, 34*, 58–73. https://doi.org/10.1016/j.learninstruc.2014.08.002

Karjanto, N., & Yong, S. T. (2013). Test anxiety in mathematics among early undergraduate students in a British university in Malaysia. *European Journal of Engineering Education, 38*(1), 11–37. https://doi.org/10.1080/03043797.2012.742867

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469–486. https://doi.org/10.1037/a0017341

Kelemen, W. L., & Weaver, C. A. III (1997). Enhanced memory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(6), 1394–1409. https://doi.org/10.1037/0278-7393.23.6.1394

Kimball, D. R., Smith, T. A., & Muntean, W. J. (2012). Does delaying judgments of learning really improve the efficacy of study decisions? Not so much. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 923–954.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2012). The relationships between monitoring, regulation and performance. *Learning and Instruction, 22*(4), 296–298. https://doi.org/10.1016/j.learninstruc.2012.01.002

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*(6), 585–592. https://doi.org/10.1111/j.1467-9280. 2008.02127.x

Kornell, N., & Finn, B. (2016). Self-regulated learning: An overview of theory and data. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 325–340). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.23

Kornell, N., & Rhodes, M. G. (2013). Feedback reduces the metacognitive benefit of tests. *Journal of Experimental Psychology: Applied, 19*(1), 1–13. https://doi.org/10.1037/a0032147

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15*(4), 307–318. https://doi.org/10.1037/a0017599

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*(2), 152–166. https://doi.org/10.1016/j.jecp.2008.10.002

Little, J. L., & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition, 43*(2), 283–297. https://doi.org/10.3758/s13421-014-0475-1

Maki, R. H. (1998). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition, 26*(5), 959–964. https://doi.org/10.3758/BF03201176

Masson, M. E., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 509–527. https://doi.org/10.1037/a0014876

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20*(4), 516–522. https://doi.org/10.1111/j.1467-9280.2009.02325.x

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*(3), 159–163. https://doi.org/10.1111/j.1467-8721.2009.01628.x

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition, 29*, 131–140. https://doi.org/10.1016/j.concog.2014.08.008

Miller, T. M., & Geraci, L. (2016). The influence of retrieval practice on metacognition: The contribution of analytic and non-analytic processes. *Consciousness and Cognition, 42*, 41–50. https://doi.org/10.1016/j.concog.2016.03.010

Mitsuda, M. (1988). Effects of processing variables on comprehension monitoring activities. *Psychologia: An International Journal of Psychology in the Orient, 31*(2), 98–105.

Mudrick, N. V., Sawyer, R., Price, M. J., Lester, J., Roberts, C., & Azevedo, R. (2018). Identifying how metacognitive judgments influence student performance during learning with MetaTutorIVH. In R. Nkambou, R. Azevedo, & J. Vassileva (Eds.), *International Conference on Intelligent Tutoring Systems* (pp. 140–149). Springer. https://doi.org/10.1007/978-3-319-91464-0_14

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychology Bulletin, 95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*(2), 102–116. https://doi.org/10.1037/0003-066X.51.2.102

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation, 26*, 125–141. https://doi.org/10.1016/S0079-7421(08)60053-5

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Educational, 74*(1), 7–28.

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*(2), 159–179. https://doi.org/10.1007/s10409-006-9595-6

Prinz, A., Golke, S., & Wittwer, J. (2018). The double curse of misconceptions: Misconceptions impair not only text comprehension but also metacomprehension in the domain of statistics. *Instructional Science, 46*(5), 723–765. https://doi.org/10.1007/s11251-018-9452-6

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*(1), 69–80. https://doi.org/10.1037/0278-7393.28.1.69

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4–5), 559–579. https://doi.org/10.1080/09541440701326022

Rawson, K. A., & Dunlosky, J. (2013). Retrieval-monitoring-feedback (RMF) technique for producing efficient and durable student learning. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 67–78). Springer. https://doi.org/10.1007/978-1-4419-5546-3_5

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*(4), 523–548. https://doi.org/10.1007/s10648-013-9240-4

Rawson, K. A., O'Neil, R., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied, 17*(3), 288–302. https://doi.org/10.1037/a0024749

Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction, 22*(4), 262–270. https://doi.org/10.1016/j.learninstruc.2011.10.007

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 65–80). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.4

Rhodes, M. G. (2019). Metacognition. *Teaching of Psychology, 46*(2), 168–175. https://doi.org/10.1093/oxfordhb/9780199336746.013.4

Rhodes, M. G., Cleary, A. M., & DeLosh, E. L. (2020). *A guide to effective studying and learning: Practical strategies from the science of learning*. Oxford University Press.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*(1), 131–148. https://doi.org/10.1037/a0021705

Roll, I., Aleven, V., McLaren, B. M., Ryu, E., Baker, R. S., & Koedinger, K. R. (2006). The help tutor: Does metacognitive feedback improve students' help-seeking actions, skills and learning? In M. Ikeda, K. D. Ashley, & T. W. Chan (Eds.), *International Conference on Intelligent Tutoring Systems* (pp. 360–369). Springer. https://doi.org/10.1007/11774303_36

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Rutten, N., Van Joolingen, W. R., & Van Der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education, 58*(1), 136–153. https://doi.org/10.1016/j.compedu.2011.07.017

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology, 33*(5), 918–929. https://doi.org/10.1002/acp.3556

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51*(1), 71–79. https://doi.org/10.1016/j.jml.2004.03.004

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory, 18*(7), 698–711. https://doi.org/10.1080/09658211.2010.506441

Sidi, Y., Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction, 51*, 61–73. https://doi.org/10.1016/j.learninstruc.2017.01.002

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition, 3*(3), 161–164. https://doi.org/ 10.1016/j.jarmac.2014.02.001

Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*(2), 129–160. https://doi.org/10.1016/S0361-476X(02)00011-5

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology, 86*(2), 290–302. https://doi.org/10.1037/0022-0663.86.2.290

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1267–1280. https://doi.org/10.1037/0278-7393.31.6.1267

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331–362. https://doi.org/10.1080/01638530902959927

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In *Handbook of metacognition in education* (pp. 97–118). Routledge.

Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2012). Elementary school experience with comprehension testing may influence metacomprehension accuracy among seventh and eighth graders. *Journal of Educational Psychology, 104*(3), 554–564. https://doi.org/10.1037/a0028660

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*(2), 264–273. https://doi.org/10.1348/135910710X510494

Thomas, R. C., Finn, B., & Jacoby, L. L. (2016). Prior experience shapes metacognitive judgments at the category level: The role of testing and category difficulty. *Metacognition and Learning, 11*(3), 257–274. https://doi.org/10.1007/s11409-015-9144-4

Thomas, A. K., & McDaniel, M. A. (2007a). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition, 35*(4), 668–678.

Thomas, A. K., & McDaniel, M. A. (2007b). Metacomprehension for educationally relevant materials: Dramatic effects of encoding-retrieval interactions. *Psychonomic Bulletin & Review, 14*(2), 212–218. https://doi.org/10.3758/BF03193305

Van Loon, M. H., de Bruin, A. B., Van Gog, T., & Van Merriënboer, J. J. (2013a). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction, 24*, 15–25. https://doi.org/10.1016/j.learninstruc.2012.08.005

Van Loon, M. H., De Bruin, A. B., Van Gog, T., & Van Merriënboer, J. J. (2013b). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning, 8*(2), 173–191. https://doi.org/10.1007/s11409-013-9100-0

Van Loon, M. H., De Bruin, A. B., Van Gog, T., Van Merriënboer, J. J., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154. https://doi.org/10.1016/j.actpsy.2014.06.007

Von der Linden, N., Löffler, E., & Schneider, W. (2017). Relation of metacognitive monitoring and control processes across the life-span. *Journal of Educational and Developmental Psychology, 7*(1), 86–101.

Walczyk, J. J., & Hall, V. C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology, 81*(3), 435–437. https://doi.org/10.1037/0022-0663.81.3.435

Wiley, J., Griffin, T. D., Jaeger, A. J., Jarosz, A. F., Cushen, P. J., & Thiede, K. W. (2016). Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied, 22*(4), 393–405.

Wiley, J., Jaeger, A. J., Taylor, A. R., & Griffin, T. D. (2018). When analogies harm: The effects of analogies on metacomprehension. *Learning and Instruction, 55*, 113–123. https://doi.org/10.1037/xap0000096

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology, 27*(4), 551–572. https://doi.org/10.1016/S0361-476X(02)00006-1

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory, 20*(6), 568–579. https://doi.org/10.1080/09658211.2012.687052

Zamary, A., Rawson, K. A., & Dunlosky, J. (2016). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Learning and Instruction, 46*, 12–20. https://doi.org/10.1016/j.learninstruc.2016.08.002

History
Received March 17, 2020
Revision received August 23, 2020
Accepted August 24, 2020
Published online April 14, 2021

Conflict of Interest
The authors report no conflicts of interest.

Funding
This work was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1321845 to the first author and Grant No. 006784-00002 to the second author.

Hannah Hausman
Psychology Department
University of California
1156 High Street
Santa Cruz, CA 95064
USA
hannah.erin.hausman@gmail.com