Journal of Educational Psychology
2012, Vol. 104, No. 4, 922–931

0022-0663/12/$12.00 DOI: 10.1037/a0028661

Selective Benefits of Question Self-Generation and Answering for Remembering Expository Text

Julie M. Bugg and Mark A. McDaniel
Washington University in St. Louis

The present study examined possible memory and metacomprehension benefits of using a combined question self-generation and answering technique, relative to rereading, as a study strategy for expository passages. In the 2 question self-generation and answering conditions (detail or conceptual questions), participants were prompted on how to generate questions of a particular type (detail or conceptual) and given practice and feedback prior to reading and studying 4 experimental passages. Participants then made judgments of learning for detailed and conceptual information from the passages, following which a cued-recall test with detail and conceptual questions was administered. The self-generation and answering of conceptual questions yielded a significant benefit to memory performance for conceptual but not detailed test questions, relative to a rereading condition, whereas the self-generation and answering of detail questions provided no benefit. A similar pattern was found for metacomprehension as assessed by calibration, but not relative monitoring accuracy. The selective memory benefit observed here is consistent with theoretical frameworks that emphasize the importance of transfer- and material-appropriate processing in modulating the benefits of using question self-generation and answering as a study strategy.

Keywords: study strategies, self-generated questions, memory, metacomprehension

In educational settings, reading assignments are a ubiquitous aspect of instruction, particularly at the college level (e.g., reading assignments are a core component of college syllabi). Theorists and educators alike would agree that effective learning requires that students engage in active, elaborative processing while reading (see, e.g., McNamara, 2004; Rosenshine, Meister, & Chapman, 1996) and that students accurately monitor their comprehension so as to effectively calibrate their degree of learning (Dunlosky & Metcalfe, 2009). Unfortunately, research suggests that some readers may not ordinarily engage processing strategies that support optimal learning. For instance, according to some theories, characteristically some readers are "lazy," creating only the minimal representational structure necessary to comprehend the text (see Fletcher & Bloom, 1988) and failing to pause to construct inferences necessary for complete understanding of technical text (Noordman, Vonk, & Kempff, 1992). Furthermore, some readers, even at the college level, appear to be relatively inaccurate at gauging how much they understand and will remember about the content of an expository passage (see Glenberg & Epstein, 1985; Maki & Berry, 1984).

This article was published Online First May 28, 2012.

Julie M. Bugg and Mark A. McDaniel, Department of Psychology, Washington University in St. Louis.

This research was supported by collaborative activity James S. McDonnell Foundation Grant 220020166. The authors are grateful to Jivata Raja, Sneha Thakur, and Jessye Brick for assistance with data collection and scoring.

Correspondence concerning this article should be addressed to Julie M. Bugg, Washington University in St. Louis, Department of Psychology, Campus Box 1125, One Brookings Drive, St. Louis, MO 63130. E-mail: jbugg@artsci.wustl.edu

Accordingly, much research in education and psychology has been directed at identifying techniques to stimulate learners to more actively and effectively process text materials. In this study, we examine one such technique, that of requiring readers to self-generate and answer questions about the content in the text. This is a technique that Mayer (2003) has advocated as a means for making the relatively passive activity of reading a more active learning experience. The idea is that self-generating and answering questions encourages readers to elaborate and consider the contents of the text more fully than they otherwise would, leading to better learning of the text material (for a review, see Wong, 1985). Experimental investigations, however, have not uniformly supported this claim, likely in part because as reviewed in Wong and Rosenshine et al. (1996), a wide range of methods have been used to stimulate self-generation of questions (and additionally, answering of these questions, in some studies).

These reviews suggest generally that methods incorporating procedural prompts (i.e., training/guiding readers in self-generating questions) are more effective in improving comprehension and retention of target content than are methods that do not incorporate procedural prompts (e.g., Davey & McBride, 1986; though see Foos, Mora, & Tkacz, 1994, for evidence that prompts are not necessary to produce benefits of question self-generation and answering). But even within the domain of procedural prompts, a number of different types of prompts have been used, and the results in the literature vary somewhat unsystematically across prompt type. This state of affairs prompted Wong (1985) to conclude that only by identifying those psychological processes that are elicited by specific types of questions (prompts) can a clear explanation of what mediates benefits of self-question generation be attained.

As a step in this direction, we examined in the present study relatively well-defined prompts that are characterized as "question types" (Rosenshine et al., 1996). Raphael and Pearson (1985) identified three question types, with two of the types in part motivating the present study. For one type, the answer is found in a single sentence, and for the second type, the answer requires integration across two or more sentences. This classification was developed to assist learners in improving their performance on criterial tests (Raphael & Pearson) and has been extended as prompts to guide readers' self-generation of questions (Dermody, 1988; Labercane & Battle, 1987; Smith, 1977). In the sole published study examining the Raphael and Pearson question-type prompts, grade-school children with learning disabilities were prompted on how to generate different question types (in addition to receiving guidance on other strategies such as summarizing) in 28 sessions of training. The training did not, however, lead to significant gains in reading achievement relative to a control condition (Labercane & Battle). In the present study, we directly contrast two question-type prompts similar to those described by Raphael and Pearson, and we do so in a single session with college students. No study has explicitly contrasted the potential benefits associated with the different question-type prompts of interest here, but one recent study hints that a single session may be sufficient for training college students to self-generate questions that can improve memory for textually presented information. Weinstein, McDermott, and Roediger (2010) gave college student participants "comprehension questions" that other students had previously generated as prompts to guide the kinds of questions that the participants should generate and answer during their own reading. This self-questioning and answering group showed better memory for the information in the texts than did a group who reread the texts. Weinstein et al. did not, however, manipulate different types of questions that could be self-generated, assess whether the benefits depended on the types of questions generated, or analyze metacomprehension accuracy; we explore these novel factors in the present experiment.

Following the Raphael and Pearson (1985) framework, we prompted different groups of participants to generate and, in addition, answer one of two types of questions to help them study target passages (i.e., expository texts). For ease of exposition and to facilitate our theoretical analysis of how these different question types might affect memory and metacomprehension, we label one type of question a *detail question* (the answer referred to a detail or fact that could be found in a single sentence) and the other type of question a *conceptual question* (the answer integrated thematic information across two or more sentences) (cf. Thomas & McDaniel, 2007). The question self-generation conditions were compared with a reread control condition. We considered the reread condition to be a strong control because it required readers to spend additional time on the texts, following an initial reading, as did the question self-generation groups. To comprehensively gauge the mnemonic outcomes of question self-generation and to provide leverage on the theoretical views developed below, we administered a final memory test that probed for both details and conceptual information. In addition, we assessed the accuracy of participants' metacomprehension of the passages—the degree to which the reader can judge his or her own learning of text materials (Dunlosky & Lipko, 2007). Metacomprehension accuracy may be an equally important outcome in educational contexts because it is assumed that accurate metacomprehension effectively guides subsequent study (i.e., students spend additional time studying passages deemed to be not well learned; for a review, see Son & Metcalfe, 2000).

Theoretical Predictions

A priori, several different patterns seemed possible. From the perspective that readers will learn more when they are encouraged to more actively process the text, the most straightforward expectation is that the question self-generation and answering groups would generally display better memory for the contents of the text than the reread control group. We believed this was not implausible, reasoning that self-generation and answering of questions, even if focused on one particular type of content (detail, conceptual), would require at the minimum that readers consciously evaluate the nature of the information they were considering as the basis for question generation (detail, conceptual). Such elaborative processing might benefit retention of both types of information even if the questions that were self-generated and answered were limited to one type of information (detail, conceptual). By the same token, metacomprehension accuracy might also be generally improved by question self-generation. This expectation is based on findings showing that techniques that stimulate more active processing of text, such as concept mapping and self-explanation, significantly improve metacomprehension accuracy (e.g., Griffin, Wiley, & Thiede, 2008; Maki, Foley, Kajer, Thompson, & Willert, 1990; Thiede, Anderson, & Therriault, 2003; Thiede, Griffin, Wiley, & Anderson, 2010).

Alternatively, the transfer-appropriate-processing approach to memory suggests that elaborative processing will benefit memory only to the extent that the processing overlaps with the information required at test (McDaniel, Friedman, & Bourne, 1978; Morris, Bransford, & Franks, 1977; Roediger, Weldon, & Challis, 1989). The transfer-appropriate-processing framework can be directly applied to the present context, generating the following predictions. Prompting self-generation and answering of detail questions should improve memory for details but not conceptual information relative to the reread control. By contrast, prompting self-generation and answering of conceptual questions should improve memory for conceptual information but not details (relative to the reread control).

Recent work has also shown that transfer-appropriate-processing dynamics appear to hold for metacomprehension (Thomas & McDaniel, 2007). Thomas and McDaniel (2007) found that metacomprehension was more accurate for detailed test questions when participants engaged in detailed processing at study (i.e., inserted missing letters in words; see also Maki et al., 1990) than when they engaged in conceptual processing at study (i.e., sorted sentences into a coherent paragraph), with the reverse being found for conceptual test questions. Although these findings are limited to study strategies that are not educationally authentic, that is, strategies not typically adopted by students or advocated by educators, they do propel the following predictions for the present study. Metacomprehension accuracy for details should be higher after self-generation and answering of detail questions, whereas metacomprehension accuracy for conceptual information should be higher after self-generation and answering of conceptual questions, relative to the reread condition.

There is yet a third, more nuanced and perhaps counterintuitive set of predictions based on a contextual framework developed to anticipate effects of more engaged processing on learning (e.g., stimulated by introducing difficulties into the learning environment; see McDaniel & Butler, 2011, for a complete presentation of this framework and related evidence). Briefly, this framework assumes that in addition to the transfer appropriateness of processing, one must consider the processing normally invited by the materials themselves (termed *material appropriate processing*; Einstein, McDaniel, Owen, & Cote, 1990; McDaniel & Einstein, 1989, 2005). Prior research suggests that different types of text invite encoding and retention of somewhat different types of information, with well-structured narratives (such as folktales) inviting processing of conceptual information that interrelates propositions in the text (McDaniel, Hines, Waddill, & Einstein, 1994) and expository text inviting more focus on details (see McDaniel & Einstein, 1989, for a more extended discussion of this assumption). Furthermore, the effectiveness of any particular study strategy is anticipated to depend on the degree to which the study strategy stimulates processing of information that is not normally invited by the particular target passages.

For present purposes, the key point regards study strategies that are anticipated to be effective for expository text. Because expository text invites more focus on details (perhaps because there is an absence of a ready conceptual structure), strategies that stimulate processing of relational information (i.e., information that integrates across several sentences or propositions) will be complementary to the processing already invited by expository passages, and consequently will benefit memory (recall). However, study strategies that stimulate processing of details (or what has been termed "proposition-specific" elaboration) will be relatively redundant with that normally invited by expository text, and consequently have little benefit on memory (see Einstein et al., 1990, for supporting experiments; see McDaniel & Einstein, 2005, for a review). Applied to the present question self-generation and answering conditions, this framework anticipates that the self-generation and answering of conceptual questions will improve memory performance on final conceptual questions relative to the reread control. More provocatively, this framework anticipates that the focus on detailed information presumably prompted by self-generating and answering detailed questions will produce no significant improvement for the detail questions on the final test relative to the reread condition (because the prompted focus on details is assumed to be relatively redundant with that invited by expository texts). With regard to metacomprehension accuracy, we thought it possible that the just predicted memory patterns would extend to metacomprehension accuracy. Specifically, self-generation and answering of conceptual questions might selectively enhance metacomprehension accuracy for conceptual questions.

Method

Participants

Forty-eight undergraduates participated for course credit or were paid at a rate of $10/hr. Sixteen participants were randomly assigned to each of the three study strategy conditions.

Design and Procedure

A 3 × 2 mixed subjects design was used, with study strategy as a between-subjects factor and type of test question as a within-subjects factor. The three study strategy conditions were reread, self-generation and answering of detailed questions, and self-generation and answering of conceptual questions. During study, participants in the reread condition read the entire passage once, and then read it a second time. Participants in the detailed and conceptual question generation and answering conditions read the entire passage once and then generated either detail questions or conceptual questions (and corresponding answers), respectively. During test, detail and conceptual test questions were administered to all participants. A *detail question* was defined as one that referred to a detail or fact contained within a single sentence in the passage. A *conceptual question* referred to the overall gist or a major theme and required integration of information across (at least two) sentences.

Following informed consent, participants were instructed that they would be reading and studying several passages for a later test. In the reread condition, participants read the entire passage once, informed the experimenter when they were done, and were then asked to study the passage by rereading it. Participants in the question generation and answering conditions were instructed to read the entire passage and inform the experimenter when they were done. They were then asked to study the passage according to the following prompt:

> When studying the passages, we would like you to use a unique strategy in which you imagine that you are a teacher who has assigned the passage to his/her class to read. You, as the teacher, are now developing a test to administer to your students to assess their knowledge of information from the passage. For each passage, you will be asked to generate 3 questions and their corresponding answers.

Participants had access to the passage while generating and answering the questions (i.e., while studying). Participants were told that the questions should assess students' knowledge of a detail or fact from a single sentence in the passage (detail question generation condition) or knowledge of concepts or information that must be integrated across sentences in the passage (conceptual question generation condition).

Following the initial set of instructions, an excerpt from a sample passage was shown. Participants in the reread condition simply read the sample passage, whereas those in the question-generation conditions read the passage and were shown a sample detail question or a sample conceptual question (depending on the question-generation condition assigned). The experimenter explained why the question was designated as a particular type (detail or conceptual). Next, a practice passage was given. Participants in the reread condition read, and then reread the passage. Those in the question-generation conditions were asked to read the passage and engage in the study task. For the practice passage, participants were prompted to generate only two (detail or conceptual) questions. Participants wrote the questions and the corresponding answers on a response sheet. The experimenter then provided the participant with individualized feedback pertaining to the questions they had generated. They were told whether or not the questions were of the type (detail or conceptual) requested and, if not, why they were not and how the questions could be changed to conform to the guidelines provided. Finally, the experimenter provided two sample questions of the type requested and confirmed that the participant understood the study instructions.

The experimental phase consisted of four passages. One passage was presented at a time, with the next passage provided by the experimenter after the participant read and studied the previous passage. The passages were given in the same order to all participants. Participants (except those in the reread condition) recorded self-generated questions and answers on a different response sheet for each passage. After participants read and studied all four passages, they were asked to make judgments of learning. Specifically, they were asked to judge how well they would remember the information that was contained in each of the passages using a scale ranging from 0 (*extremely unlikely to remember*) to 100 (*extremely likely to remember*). Participants were encouraged to use the full range of the scale when making their judgments. Participants provided separate judgments for each passage, beginning with the first passage they studied and moving in order to the last. Importantly, for each passage, one judgment was made for details/facts, and a second judgment was made for concepts/integrative information. Participants in the detail question-generation and answering condition made the detail/fact judgment first, and then made the concept/integrative information judgment. The reverse was true for participants in the conceptual question-generation and answering condition. The order of these judgments was counterbalanced for participants in the reread condition.

Following the metacognitive judgments, a cued-recall test was given. The test consisted of 24 questions (three detail and three conceptual questions per passage). The six test questions that corresponded to the first passage participants had read and studied were administered, followed by those corresponding to the second passage, and so on. Participants were shown the title of the passage, then the six questions that were administered in a random order. Following completion of the cued-recall test, participants were thanked and debriefed.

Materials

Following Thomas and McDaniel (2007), six expository texts from Levy (1981) were used, which were taken from the intermediate level of the Science Research Associates (SRA) reading series. The sample passage was an excerpt from the text "Skunks." The text "How Autumn Colors are Formed" was used as the practice passage. The experimental passages, in the order shown to participants, were the texts "Kanchenjunga: A Very Dangerous Mountain Range"; "Nomads of the Desert"; "The Strange Way of Spiders"; and "The Frozen Continent." The texts ranged in length from 285 to 360 words and were presented in a single-paragraph format.

Three detail and three conceptual questions were generated for each passage for the cued-recall test. Half of these questions overlapped with those used by Thomas and McDaniel (2007), and half of the questions were newly developed. For "The Frozen Continent," an example of a detail question is: "How many square miles in size is Antarctica's great ice cap?" The answer, "six million" comes from the following single sentence from the text: "Antarctica's great ice cap alone is six million square miles in size." An example of a conceptual question is: "Give two reasons why it is impossible to create a map of the crevasses in Antarctica," with the answer "Crevasses are concealed, and they are constantly moving." This answer requires integration of information across the following contiguous sentences from the text:

> The glacier is forced to change direction, causing an enormous build-up in pressure. Many crevasses are formed by this process also. Crevasses may be a concealed peril for travelers. Crevasses are often shifted great distances from the place they were created as the glacier moves. Unconcealed crevasses can easily be avoided, but most are hidden by snow bridges. The wind builds snow bridges by packing snow across the opening of the chasm. Eventually, the bridge and the crevasse become completely hidden by drifting snow.

Results

The alpha level was set at .05. Partial eta squared (η_p^2) is reported as the measure of effect size for all significant effects.

Manipulation Checks

To confirm that participants in the generation and answering conditions adhered to the instructions they were provided, we first calculated the mean percentage of questions that were of the type requested (i.e., detail for the detail question-generation/answering condition and conceptual for the conceptual question-generation/answering condition). Adherence did not differ between the detail question-generation and answering group ($M = 98\%$, $SE = 2\%$) and the conceptual question-generation and answering group ($M = 93\%$, $SE = 4\%$), $t(30) = 1.03$, $p = .311$. Next, we calculated the mean percentage of self-generated questions that were correctly answered. Accuracy was perfect (or nearly perfect) for participants in the detail ($M = 100\%$, $SE = 0\%$) and conceptual ($M = 99\%$, $SE = 1\%$) question and answering conditions, $t(30) = 1.46$, $p = .154$. Together, these findings indicate that participants in both groups followed the instructions equally well.

Cued Recall

Two raters scored responses to the cued-recall test questions. We used a partial credit scoring procedure whereby participants were awarded 0, .5, or 1 point per question.[1] Interrater reliability was high ($r = .90$). The measure of cued-recall performance was average points per question earned (e.g., a participant who earned 1 point for six of the 12 detail questions and 0 points for the other six would have a .5 score for cued-recall performance on the detail questions), and this measure was calculated separately for the 12 detail and 12 conceptual questions.

We conducted a 3 (study strategy: reread vs. generation of detail questions vs. generation of conceptual questions) × 2 (type of test question: detail vs. conceptual) mixed subjects analysis of variance (ANOVA) for cued-recall performance. Significant main effects of study strategy, $F(2, 45) = 3.21$, $MSE = .039$, $p = .050$, $\eta_p^2 = .125$, and type of test question, $F(1, 45) = 13.68$, $MSE = .016$, $p = .001$, $\eta_p^2 = .233$, were qualified by a significant interaction between

[1] We report partial credit-based performance because this scoring procedure is more common in educational contexts. Note, however, that use of a strict (i.e., nonpartial credit) scoring procedure yields a highly similar pattern of interactive effects for both cued recall and calibration.

study strategy and type of test question, $F(2, 45) = 3.55$, $MSE = .016$, $p = .037$, $\eta_p^2 = .136$.

To test the hypotheses outlined in the introduction, we decomposed the 3 × 2 interaction by analyzing the simple effects of study strategy using one-way ANOVAs and the simple effects of question type using dependent *t* tests. Cued-recall performance for the detail questions did not differ across the three study strategy groups ($F < 1$) (see Figure 1). By contrast, cued-recall performance for the conceptual questions did vary significantly as a function of study strategy, $F(2, 45) = 5.33$, $MSE = .064$, $p = .008$, $\eta_p^2 = .191$ (see Figure 1). Bonferroni post hoc comparisons indicated that a significant benefit was observed for the group that generated and answered conceptual questions ($M = .66$, $SE = .04$) relative to the group that generated and answered detail questions ($M = .47$, $SE = .04$, $p = .012$) or the group that reread ($M = .50$, $SE = .05$, $p = .043$).

Comparisons of cued-recall performance on detail and conceptual questions for each of the three groups revealed the following pattern of results. Cued-recall performance did not differ on the detail and conceptual questions for either the group that generated and answered detail questions, $t(15) = -1.12$, $p = .283$, or the group that reread, $t(15) = -0.816$, $p = .427$. However, the group who generated and answered conceptual questions performed significantly better on the conceptual questions ($M = .66$, $SE = .04$) than on the detail questions ($M = .47$, $SE = .04$), $t(15) = -5.86$, $p < .001$.

Question Overlap and Cued Recall

We examined the percentage of questions participants in each group generated that overlapped with the questions administered on the cued-recall test, as this has been shown to influence final test performance (e.g., Frase & Schwartz, 1975). Overlap was similar for the group who generated detail questions ($M = 26\%$, $SE = 2\%$) and the group who generated conceptual questions ($M = 28\%$, $SE = 3\%$), $t(30) = -0.26$, $p = .799$. To examine whether overlap was related to cued-recall performance *within* each of the question-generation and answering conditions, we compared the cued-recall performance of participants with relatively high and low question overlap. For the conceptual question-generation condition, conceptual test performance was significantly higher for participants ($n = 7$) whose questions overlapped 33%–50% (50% was the highest) with the final test ($M = .77$, $SE = .03$) than for participants ($n = 9$) whose questions had relatively low (0%–25%) overlap ($M = .58$, $SE = .06$), $t(14) = -2.58$, $p = .022$. For those whose questions had relatively low overlap, cued-recall performance approached that observed in the reread condition ($M = .50$, $SE = .05$). For the detailed question-generation condition, there was minimal effect of the degree of overlap for detail test performance. Participants ($n = 7$) whose questions had relatively high (33%–50%; 42% was the highest) overlap performed similarly ($M = .44$, $SE = .06$) to participants ($n = 9$) whose questions had relatively low (0%–25%) overlap ($M = .40$, $SE = .04$), $t(14) = -0.53$, $p = .603$.

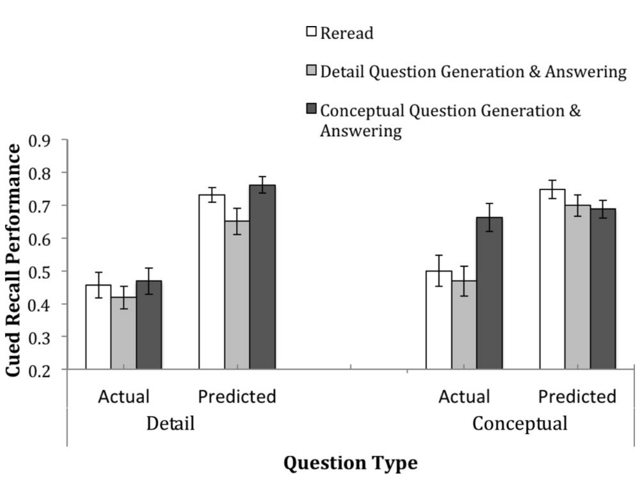

Figure 1. Actual and predicted cued-recall performance (i.e., judgments of learning) for detail and conceptual test questions as a function of study strategy. Calibration is the difference between actual and predicted performance. Error bars represent standard error.

Judgments of Learning

To examine whether judgments of learning varied as a function of study strategy and type of test question, we performed a 3 (study strategy) × 2 (type of test question) mixed ANOVA. The main effects of study strategy and question type were not significant, but the interaction of these two factors was significant, $F(2, 45) = 4.68$, $MSE = .007$, $p = .014$, $\eta_p^2 = .172$ (see Figure 1 for means). To decompose the interaction, we first analyzed the simple main effects of study strategy using one-way ANOVAs. For the detailed question type, judgments of learning did differ according to study strategy, $F(2, 45) = 3.62$, $MSE = .014$, $p = .035$, $\eta_p^2 = .139$. Bonferroni post hoc comparisons indicated that judgments of learning were significantly higher for participants who generated and answered conceptual ($M = .76$, $SE = .03$) as compared with detailed ($M = .65$, $SE = .03$) test questions ($p = .038$). For the conceptual question type, the effect of study strategy was not significant, $F(2, 45) = 1.12$, $p = .335$.

We then analyzed the simple main effects of question type using dependent *t* tests. For the group who generated and answered detail questions, judgments of learning were nominally but not significantly higher for the conceptual ($M = .70$, $SE = .03$) as compared with detailed ($M = .65$, $SE = .04$) test questions, $t(15) = -1.64$, $p = .123$. By contrast, for the group who generated and answered conceptual questions, judgments of learning were lower for the conceptual ($M = .69$, $SE = .03$) as compared with detail ($M = .76$, $SE = .02$) test questions, $t(15) = 3.20$, $p = .006$. For the reread group, judgments of learning were equivalent for conceptual ($M = .75$, $SE = .03$) and detailed ($M = .73$, $SE = .02$) test questions ($t < 1$).

Calibration

To examine metacomprehension accuracy, we derived a measure of calibration by subtracting actual cued-recall performance from predicted cued-recall performance (i.e., judgments of learning) (Dunlosky & Metcalfe, 2009). Positive values for calibration reflect overconfidence, whereas negative values reflect underconfidence. We conducted a 3 (study strategy) × 2 (type of test question) mixed ANOVA for these calibration values. A main effect of question type, $F(1, 45) = 11.13$, $MSE = .021$, $p = .002$,

$\eta_p^2 = .198$, was qualified by a significant Study Strategy × Type of Test Question interaction, $F(2, 45) = 8.12$, $MSE = .021$, $p = .001$, $\eta_p^2 = .265$.

To decompose this interaction, we conducted the analyses of simple effects as described above. Mirroring the patterns obtained for cued-recall performance, the three study strategy groups did not differ in calibration for the detail test questions ($F < 1$), but did differ in calibration for the conceptual test questions, $F(2, 45) = 8.64$, $MSE = .056$, $p = .001$, $\eta_p^2 = .277$ (see Figure 1). Bonferroni post hoc comparisons indicated that participants who generated and answered conceptual questions at study had significantly better calibration for the conceptual test questions than participants in the detail question-generation group ($p = .004$) or reread group ($p = .002$); the latter two groups did not differ ($p > .999$).

Turning to the simple effects of question type, calibration was equivalent for detailed and conceptual test questions for both the group who generated and answered detail questions, $t(15) = 0.025$, $p = .980$, and the group who reread, $t(15) = 0.425$, $p = .677$. In contrast, the group who generated and answered conceptual questions had better calibration on the conceptual test questions ($M = .03$, $SE = .04$) than on the detailed test questions ($M = .29$, $SE = .04$), $t(15) = 6.62$, $p < .001$. In fact, for the conceptual test questions, participants who generated and answered conceptual questions had calibration scores that did not differ from zero, $t(15) = 0.61$, $p = .55$, suggesting a very high degree of metacomprehension accuracy. Calibration did differ significantly from zero for all other combinations of question type and study strategy (ts > 5.36, ps $< .001$).

Relative Monitoring Accuracy

A second measure of metacomprehension accuracy is relative monitoring accuracy. Following Griffin et al. (2008), we computed Pearson correlations between each individual's cued-recall performance and their predicted performance across the four texts,[2] save for two participants in the detail question-generation group who produced no variability in predicted performance across the texts. As shown in Table 1, the average correlations were very weak for both detailed and conceptual test performance, with none differing from zero (ps $> .18$).

Analysis of the Quality of Generated Questions

Finally, we examined the quality of the generated questions by categorizing each as either a low-quality (a question requiring a verbatim restatement of information from the text) or high-quality (a question requiring inference or involving a macrostatement) question.[3] For example, a low-quality question pertaining to the excerpt on crevasses from the "Frozen Continent" text (see the Method section) is "Why are crevasses in the ice formed and often hidden?" whereas a high-quality question is "Why are crevasses dangerous formations for Antarctic travelers?" We awarded 1 point for each low-quality question and 2 points for each high-quality question such that a participant minimally earned 12 points and maximally earned 24 points. On average, the group who generated conceptual questions earned 16.5 points ($SE = .45$), whereas the group who generated detail questions earned 12.3 points ($SE = .15$), a statistically significant difference, $t(30) = -8.71$, $p < .001$. As expected, there was little variation within the detail question-generation group (range = 12–14). However, there was more variation within the conceptual question-generation group (range = 14–20), suggesting that some participants produced primarily questions that required a verbatim restatement of information, whereas others produced a mixture, including some questions that required inferencing or involved macrostatements. Consequently, we thought it informative to examine the correlations within the conceptual question-generation group between question quality, memory (cued-recall performance), and metacomprehension (calibration). There were no significant correlations between quality and memory, or quality and metacomprehension (largest $r = .17$, $p = .53$).

Table 1
Average Relative Monitoring Accuracy (Standard Error in Parentheses) as a Function of Study Strategy and Type of Test Question

	Study strategy		
Type of test	Reread	Detail question generation and answering	Conceptual question generation and answering
Detail	0.20 (.14)	−0.05 (.15)	−0.07 (.18)
Conceptual	0.04 (.15)	0.10 (.14)	−0.16 (.13)

Discussion

In the present study, we examined the effects of a combined question self-generation and answering technique on memory and metacomprehension outcomes. An overarching view has been that self-generation of questions produces more active reading, including a greater focus of attention on and elaboration of content and greater self-awareness of the degree of comprehension (Rosenshine et al., 1996), thereby leading to increased comprehension, memory, and metacomprehension. Rosenshine et al.'s review highlighted procedural prompts to guide question self-generation as an important factor influencing the attainment of such benefits. However, evidence on the effectiveness of question self-generation when accompanied by a procedural prompt is sparse with college students, with some of the primary reports examining comprehension of lectures (King, 1989, 1992) and not texts. An exception is a recent study by Weinstein et al. (2010); like Weinstein et al., we used a self-generation and answering condition and found support for the overarching view. Importantly, however, in the present study, support was qualified by a consideration of (a) the particular types of questions and answers prompted for generation, a factor not considered in previous studies in which the effects of question generation on memory and metacomprehension of text have been examined and (b) the particular types of information targeted in the test questions. Specifically, we found that generation and answering of conceptual but not detailed questions benefited performance

[2] The pattern of results was identical when gamma correlations were computed.

[3] We thank several anonymous reviewers for suggesting that we analyze the quality of the generated questions along these dimensions and examine possible relationships between question quality, memory, and metacomprehension within the conceptual question-generation condition.

on the final memory test and appeared to improve calibration relative to rereading and that these benefits were specific to conceptual test questions. Importantly, this finding cannot be explained by differential adherence to question-generation instructions across conditions, differential accuracy in answering self-generated questions, or differential degrees of overlap between the questions generated by participants and those that appeared on the final test, as these measures did not differ between the question-generation and answering groups.

This pattern suggests limits to overly general claims regarding the benefits of self-generating and answering questions as a study strategy and points to a more refined understanding of the effects of self-generation and answering of questions on memory and metacomprehension. Regarding memory, of note are two prominent components of our findings. The first is that generating and answering conceptual questions did not produce benefits to performance on the detailed questions. This finding counters the possibility outlined in the introduction that generating and answering certain types of questions, such as conceptual questions, would prompt learners to sift through both detail and conceptual (thematic) information, thereby potentially elaborating both kinds of information. Instead, the effect of generating and answering conceptual questions displayed a transfer-appropriate pattern such that the processing engaged when generating conceptual questions benefited performance on final test questions that targeted a similar level of information (i.e., conceptual). This conclusion might be further refined in light of the analysis of the quality of the generated questions. As might be expected, the conceptual question-generation and answering group generated higher quality questions (in terms of requiring inferencing and construction of macrostatements) than the detail question-generation and answering group. However, within the conceptual question-generation and answering group, generating higher quality conceptual questions was not associated with higher recall levels. Accordingly, it may be that the critical conceptual processing feature for enhancing performance on the conceptual test questions was attempting to generate and answer questions that required integration of information across sentences, as specified in the instructions.

The second component is that the just-mentioned transfer-appropriate-processing effect was observed exclusively for the generation and answering of conceptual questions. That is, a benefit to cued-recall performance on detailed questions was not observed for participants who generated and answered detailed questions at study. This potentially curious finding is readily interpreted within the material-appropriate-processing (MAP) framework outlined in the introduction. The MAP framework suggests that different types of text typically invite readers to extract particular types of information (Einstein et al., 1990; McDaniel, Einstein, Dunay, & Cobb, 1986). For present purposes, the key assumption of this framework is that individuals tend to focus on details when reading expository texts (presumably because there is no obvious organizational structure within which to integrate the content; see McDaniel & Einstein, 1989, for amplification, and Einstein et al., 1990, for supporting findings with some of the texts used in the present study). Applied to the present pattern, the framework suggests that processing stimulated by the generation and answering of detailed questions is redundant with that invited by the materials themselves, and therefore detailed questions do not provide benefits relative to reading alone (for the types of texts used here). By contrast, the processing stimulated by the generation and answering of conceptual questions is not redundant with that invited by the expository texts and accordingly, as we observed, enhances memory for this type of information.

It is worth emphasizing that the average overlap in content targeted by the self-generated conceptual questions and the criterial test conceptual questions was not particularly high (28%), suggesting the pedagogical utility of stimulating learners to attend to levels of information that expository texts (at least for the corpus used here) do not ordinarily invite learners to encode well. Even so, one would expect that the conceptual question prompts would have increasing benefits to the degree that learners could generate questions overlapping with the criterial test. This expectation was borne out in the present experiment: Conceptual test performance rose to .77 for participants whose generated conceptual questions overlapped relatively well with the final test, compared with .58 for participants with relatively low overlap.[4] Note that this pattern corresponds to the transfer-appropriate-processing interpretation discussed above; for those participants who generated conceptual questions that did not overlap with the final test (i.e., relatively low overlap in processing of question generation and final test), cued-recall performance approached that observed in the reread condition (.50).

For the detailed question-generation and answering condition, there was minimal effect of the degree of overlap for detail test performance. This pattern reinforces the conclusion that minimal benefits may be associated with self-generation and answering of questions focusing on levels of information that are redundant with that afforded by the text (during normal reading). Of course, it remains possible that if learners could somehow generate and answer questions that targeted tested details with a high degree of overlap, then self-generation and answering of detailed questions could increase learning and memory from expository text (cf. Maki et al., 1990; Thomas & McDaniel, 2007, for related findings).

For metacomprehension accuracy, as assessed by calibration, we found a similar pattern to that which was observed for memory performance. Specifically, a selective benefit was obtained for metacomprehension accuracy on conceptual test questions following the self-generation and answering of conceptual questions. This benefit reflected that the typical pattern of overconfidence in metacomprehension, observed in all other conditions in the present study, was eliminated, and high levels of metacomprehension accuracy (i.e., calibration) were obtained when the type of self-generated questions both matched the criterial final test processing and were appropriate for (i.e., not redundant with) the materials. To the extent that metacomprehension is used to guide subsequent study, elimination of overconfidence is a finding of practical importance as it anticipates more appropriate decisions regarding

[4] The fact that performance was at 77% for the highest levels of overlap may counter practical concerns regarding somewhat modest overall absolute performance levels (both for the reread and the question-generation groups). That is, in an educational setting, the overall performance levels could be viewed as unsatisfactory. For purposes of experimental design, we implemented conditions (e.g., limited exposure to the material, use of materials for which participants had little background) to obtain performance levels that would allow sensitivity to differences among experimental conditions. Notably, even under these conditions, participants who tended to generate conceptual questions that overlapped with those on the final test showed relatively high performance (77%).

study (cf. Thomas & McDaniel, 2007). There is, however, reason to be cautious in interpreting this finding; the enhanced calibration appears to be more reflective of fluctuations in cued-recall performance than it is of enhanced metacognitive monitoring (see, e.g., Connor, Dunlosky, & Hertzog, 1997; Griffin, Jee, & Wiley, 2009). That is, for the group that generated and answered conceptual questions, cued-recall performance increased for conceptual questions. As such, the generally high judgment of learning that this group assigned to conceptual questions approached actual cued-recall performance for this question type.

As for relative monitoring accuracy, the correlations between predicted performance and actual performance for the four passages were weak, with none differing from zero. It is possible that the low levels of relative monitoring accuracy in the present study may relate to the small number of passages. However, Griffin et al. (2008) used only four passages but found that relative monitoring accuracy was significant (.63) for a different type of active-processing strategy (self-explanation). A potentially important difference between the present study and that of Griffin et al. is that they obtained judgments immediately after participants read each text (see also Thomas & McDaniel, 2007), whereas we used a delayed procedure for collecting metacomprehension judgments. Some early studies suggested that relative monitoring accuracy differs from zero only when judgments and tests are immediate (Maki, 1998; see also Glenberg & Epstein, 1985, Experiment 2). However, more recent work has shown that delayed judgment tasks can improve relative monitoring accuracy. For example, it is improved when the study task (e.g., keywording, summarizing, question generation) that precedes the judgments of learning is not performed until after a delay following the completion of reading (Thiede, Dunlosky, Griffin, & Wiley, 2005; Thiede et al., 2010). The theoretical interpretation is that delayed study tasks permit access to more valid, situation-level cues that are then used as a basis for the subsequent judgments (see Thiede, Griffin, Wiley, & Redford, 2009). Accordingly, the absence of a benefit for relative monitoring accuracy in the present study may reflect the use of a delayed judgment procedure in conjunction with a study task performed immediately upon completion of reading.

To take stock, our results suggest that the benefits of self-generating and answering questions on memory for text are at least in part determined by the constellation of types of questions generated (prompted for generation) and answered, the nature of the texts, and the information targeted by the criterial test. The present results emphasize the idea that simply focusing students' attention on content by prompting generation and answering of questions is not sufficient to produce benefits over rereading alone. Paralleling basic memory research (e.g., Einstein et al., 1990; McDaniel et al., 1986; McDaniel, Einstein, & Lollis, 1988), self-generation and answering of questions appear to be beneficial primarily when attention is focused on processing information that would not ordinarily be encoded when reading particular texts (materials), and when a transfer-appropriate criterial test is administered.

Note that our interpretation has interesting but as yet untested implications with regard to extending the present findings. First, we are not arguing that generation and answering of conceptual questions is always better. For instance, for texts that do not ordinarily stimulate processing of details (e.g., a folktale), generating conceptual questions may not yield significant benefits. Instead, generation of detailed questions may serve learners better in improving memory and metacomprehension accuracy for this genre of text (see Einstein et al., 1990). Admittedly, such texts (e.g., folktales) are rarely found in educational settings. Still, in educational settings, learners with expertise in a content area might easily organize expository texts from that content area (cf. McNamara, Kintsch, Butler-Songer, & Kintsch, 1996) and consequently could benefit more from detailed question generation than conceptual question generation. A test of these predictions, in concert with the present data, would fully inform the material appropriate processing interpretation of the present patterns.

A more straightforward potential implication of the present findings concerns limitations to the value of self-generation and answering of questions as a study strategy in educational settings. Self-generation and answering of questions takes additional time over rereading (e.g., Weinstein et al., 2010). Consequently, unless educators are alert to the demands of the criterial test and processing afforded by materials themselves, encouraging students to engage in question generation and answering could prove to be labor in vain on the student's part. Specifically, for educational settings in which students are studying expository text, our findings suggest that prompting self-question generation and answering may not be optimal when criterial tests focus on details or when students are prompted to generate detail questions. We offer these implications cautiously because learner ability, prior knowledge, or even length of the texts (i.e., we used relatively short passages) could further qualify the present patterns.

In closing, we refer to Rosenshine et al.'s (1996) lament "at the present time developing procedural prompts appears to be an art" (p. 198). Fifteen years later, not much has changed, as very little empirical work has been published to evaluate the benefits of different procedural prompts (i.e., prompts to stimulate generation of different question types). The present experiment thus provides needed empirical evaluation of two previously identified question types (Raphael & Pearson, 1985), and importantly indicates that the benefits are nuanced. Within a single experiment, we observed a mixed pattern such that the benefits of question generation and answering on memory and metacomprehension were selective. Following appeals in the literature (Rosenshine et al.; Wong, 1985), we have proposed a contextual approach drawn from basic memory theory (Jenkins, 1979; McDaniel & Butler, 2011) in an attempt to forge a coherent understanding of how self-questioning and answering works—when it is beneficial and when it is not. We do not intend to imply that the present experimental findings confirm this approach but rather that they hint at the fruitfulness of simultaneously considering transfer and material appropriate processing dynamics on the benefits of question self-generation and answering for improving memory and metacomprehension of text.

References

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12,* 50–71. doi:10.1037/0882-7974.12.1.50

Davey, B., & McBride, S. (1986). Effects of question-generation training on reading comprehension. *Journal of Educational Psychology, 78,* 256–262. doi:10.1037/0022-0663.78.4.256

Dermody, M. (1988, February). *Metacognitive strategies for development of reading comprehension for younger children.* Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, New Orleans, LA.

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

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Thousand Oaks, CA: Sage Publications.

Einstein, G. O., McDaniel, M. A., Owen, P. D., & Cote, N. C. (1990). Encoding and recall of texts: The importance of material appropriate processing. *Journal of Memory and Language, 29,* 566–581. doi:10.1016/0749-596X(90)90052-2

Fletcher, C. R., & Bloom, C. P. (1988). Causal reasoning in the comprehension of simple narrative texts. *Journal of Memory and Language, 244,* 235–244.

Foos, P. W., Mora, J. J., & Tkacz, S. (1994). Student study techniques and the generation effect. *Journal of Educational Psychology, 86,* 567–576. doi:10.1037/0022-0663.86.4.567

Frase, L. T., & Schwartz, B. J. (1975). Effect of question production and answering on prose recall. *Journal of Educational Psychology, 67,* 628–635. doi:10.1037/0022-0663.67.5.628

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

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37,* 1001–1013. doi:10.3758/MC.37.7.1001

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,* 93–103. doi:10.3758/MC.36.1.93

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 429–446). Hillsdale, NJ: Erlbaum.

King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology, 14,* 1–16.

King, A. (1992). Facilitating elaborative learning through guided student-generated questioning. *Educational Psychologist, 27,* 111–126. doi:10.1207/s15326985ep2701_8

Labercane, G., & Battle, J. (1987). Cognitive processing strategies, self-esteem, and reading comprehension of learning disabled students. *B. C. Journal of Special Education, 11,* 167–185.

Levy, B. A. (1981). Interactive processes during reading. In. A. M. Lesgold & C. A. Perfetti (Eds.), *Interactive processes in reading* (pp. 1–35). Hillsdale, NJ: Erlbaum.

Maki, R. H. (1998). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition, 26,* 959–964.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 663–679. doi:10.1037/0278-7393.10.4.663

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 609–616. doi:10.1037/0278-7393.16.4.609

Mayer, R. E. (2003). *Learning and instruction.* Upper Saddle River, NJ: Merrill Prentice Hall.

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. In A. Benjamin (Ed.), *Successful remembering and successful Forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–199). New York, NY: Taylor & Francis.

McDaniel, M. A., & Einstein, G. O. (1989). Material appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review, 1,* 113–145. doi:10.1007/BF01326639

McDaniel, M. A., & Einstein, G. O. (2005). Material appropriate difficulty: A framework for determining when difficulty is desirable for improving learning. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 73–85). Washington, DC: American Psychological Association. doi:10.1037/10895-006

McDaniel, M. A., Einstein, G. O., Dunay, P. K., & Cobb, R. (1986). Encoding difficulty and memory: Toward a unifying theory. *Journal of Memory and Language, 25,* 645–656. doi:10.1016/0749-596X(86)90041-0

McDaniel, M. A., Einstein, G. O., & Lollis, T. (1988). Qualitative and quantitative considerations in encoding difficulty effects. *Memory & Cognition, 16,* 8–14. doi:10.3758/BF03197740

McDaniel, M. A., Friedman, A., & Bourne, L. E. (1978). Remembering the levels of information in words. *Memory & Cognition, 6,* 156–164. doi:10.3758/BF03197441

McDaniel, M. A., Hines, R. J., Waddill, P. J., & Einstein, G. O. (1994). What makes folk tales unique: Content familiarity, causal structure, scripts, or superstructures? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 169–184. doi:10.1037/0278-7393.20.1.169

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38,* 1–30. doi:10.1207/s15326950dp3801_1

McNamara, D. S., Kintsch, E., Butler-Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43. doi:10.1207/s1532690xci1401_1

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 519–533. doi:10.1016/S0022-5371(77)80016-9

Noordman, L. G. M., Vonk, W., & Kempff, H. J. (1992). Causal inferences during the reading of expository texts. *Journal of Memory and Language, 31,* 573–590. doi:10.1016/0749-596X(92)90029-W

Raphael, T. E., & Pearson, P. D. (1985). Increasing students' awareness of sources of information for answering questions. *American Educational Research Journal, 22,* 217–235.

Roediger, H. L., III, Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 3–41). Hillsdale, NJ: Erlbaum.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research, 66,* 181–221.

Smith, N. J. (1977). *The effects of training teachers to teach students at different reading ability levels to formulate three types of questions on reading comprehension and question generation ability* (Unpublished doctoral dissertation). University of Georgia.

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. doi:10.1037/0278-7393.26.1.204

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

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,* 1267–1280. doi:10.1037/0278-7393.31.6.1267

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

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85–106). New York, NY: Routledge.

Thomas, A. K., & McDaniel, M. A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition, 35,* 668–678. doi:10.3758/BF03193305

Weinstein, Y., McDermott, K. B., & Roediger, H. L. (2010). A comparison of study strategies for passages: Re-reading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied, 16,* 308–316. doi:10.1037/a0020992

Wong, B. Y. L. (1985). Self-questioning instructional research: A review. *Review of Educational Research, 55,* 227–268.

Received March 6, 2011
Revision received March 27, 2012
Accepted April 10, 2012 ■