Journal of Educational Psychology

Guided Retrieval Practice of Educational Materials Using Automated Scoring

Phillip J. Grimaldi and Jeffrey D. Karpicke

Online First Publication, June 24, 2013. doi: 10.1037/a0033208

CITATION

Grimaldi, P. J., & Karpicke, J. D. (2013, June 24). Guided Retrieval Practice of Educational Materials Using Automated Scoring. *Journal of Educational Psychology*. Advance online publication. doi: 10.1037/a0033208

Journal of Educational Psychology

© 2013 American Psychological Association
0022-0663/13/$12.00 DOI: 10.1037/a0033208

Guided Retrieval Practice of Educational Materials Using Automated Scoring

Phillip J. Grimaldi and Jeffrey D. Karpicke
Purdue University

Retrieval practice is a powerful way to promote long-term retention and meaningful learning. However, students do not frequently practice retrieval on their own, and when they do, they have difficulty evaluating the correctness of their responses and making effective study choices. To address these problems, we have developed a guided retrieval practice program that uses an automated scoring algorithm, called QuickScore, to evaluate responses during retrieval practice and make study choices based on student performance. In Experiments 1A and 1B, students learned human anatomy materials in either repeated retrieval or repeated study conditions. Repeated retrieval in the computer-based program produced large gains in retention on a delayed test. In Experiment 2, we examined the accuracy of QuickScore's scoring relative to students' self-scoring of their own responses. Students exhibited a dramatic bias to give partial or full credit to completely incorrect responses, while QuickScore was far less likely to score incorrect responses as correct. These results support the efficacy of computer guided retrieval practice for promoting long-term learning.

Keywords: retrieval based learning, computer based learning, testing effect

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

In recent years, there has been a surge of research demonstrating that retrieval practice produces powerful benefits for learning (Karpicke, 2012; Karpicke & Grimaldi, 2012; Roediger & Butler, 2011). Several studies have shown that repeated retrieval enhances long-term retention of relatively simple materials like lists of words or paired-associates (e.g., Karpicke & Roediger, 2007, 2008; Pyc & Rawson, 2009, 2010). Recent research has shown that practicing retrieval enhances the learning of more complex and educationally relevant text materials (e.g., Karpicke & Roediger, 2010; McDaniel, Howard, & Einstein, 2009; Roediger & Karpicke, 2006; Wissman, Rawson, & Pyc, 2011) and improves performance on assessments that include conceptual and inferential questions (Butler, 2010; Johnson & Mayer, 2009; Karpicke & Blunt, 2011). Thus, retrieval practice is a robust and reliable strategy for enhancing meaningful learning.

A current challenge is to identify the best ways to leverage retrieval practice within educational activities. One approach is to introduce more low stakes quizzing into the classroom, and several recent studies have demonstrated positive effects of classroom quizzing on long-term retention (Mayer et al., 2009; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; Roediger, Agarwal, McDaniel, & McDermott, 2011). A second approach is to encourage students to practice retrieval outside the classroom, an approach that would not consume valuable class time and should improve students' preparation for classroom activities. However, the effectiveness of retrieval practice outside the classroom weighs heavily on students' abilities to monitor and regulate their own learning. Students would need to recognize that retrieval practice is an effective learning strategy; they would need to choose to repeatedly retrieve material in the most effective ways; and they would need to correctly evaluate when their retrieval attempts have been successful. There is now good evidence that students struggle with all of these metacognitive abilities (Dunlosky & Rawson, 2012; Karpicke, 2009; Karpicke, Butler, & Roediger, 2009; Kornell & Son, 2009). The present experiments support the development of a computer-based retrieval practice program that automatically scores students' responses and guides them to practice retrieval in effective ways. Before describing the three experiments, we first provide an overview of the difficulties students often have in monitoring and regulating their own retrieval practice.

Students Lack Awareness of the Benefits of Retrieval Practice

If students were aware that the act of retrieving knowledge produced learning, then their metacognitive judgments of learning would reflect that students thought they learned more after practicing retrieval than after engaging in other study conditions. A consistent finding in research on retrieval practice is that most students lack metacognitive awareness of the benefits of practicing

,

Phillip J. Grimaldi and Jeffrey D. Karpicke, Department of Psychological Sciences, Purdue University.

This research and the writing of this article were supported in part by National Science Foundation Grants DUE-0941170 and DRL-1149363 and by the Institute of Education Sciences, U.S. Department of Education Grant R305A110903. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We thank Lauren Abplanalp, Corey Guilford, Haleigh Hoffer, Stephanie Karpicke, Breanne Lawler, Sohail Mehra, Laurel Poston, Michelle Schwager, and Caitlyn Taylor for their help in data collection and scoring. We also thank Lisa Hilliard for her help in creating the materials.

Correspondence concerning this article should be addressed to Phillip J. Grimaldi, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2081. E-mail: pgrimald@purdue.edu

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

retrieval (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Karpicke, 2009; Karpicke & Blunt, 2011; Karpicke et al., 2009; Karpicke & Roediger, 2008; Kornell & Son, 2009; Roediger & Karpicke, 2006). In these experiments on metacognitive monitoring, students repeatedly read or repeatedly recalled material and then were asked to predict their performance on a future test. Students who practiced repeated retrieval consistently predicted lower performance than students who repeatedly studied or engaged in other activities (e.g., Karpicke & Blunt, 2011; Roediger & Karpicke, 2006). Although practicing retrieval often produces substantial benefits for long-term learning, many students are unaware that this is true.

Students Do Not Choose to Practice Repeated Retrieval

It follows that if students are not aware that retrieval practice enhances learning, they will be unlikely to practice retrieval when they regulate their own learning. Laboratory studies and surveys of students' real-world study behaviors have confirmed that many students do not choose to practice retrieval under circumstances in which they control and regulate their own learning. For example, Karpicke (2009) had students learn foreign language word pairs across a series of alternating study and recall trials. Once the students had successfully recalled an item, they were given three options about what do with it: They could remove it from further practice, restudy it two more times, or practice retrieving it two more times. Practicing retrieval, even just two additional times for each item, produced large gains in long-term retention (see too Karpicke & Roediger, 2007, 2008; Karpicke & Smith, 2012). Yet when the students were given control over their own learning, they overwhelmingly chose to remove what they had recalled (60% of the items) rather than practice repeated retrieval (25% of the items). Thus, when students regulate their own learning, they often practice to the criterion of one correct recall of each item. Repeated retrieval practice would produce large gains in learning (see too Rawson & Dunlosky, 2011), but students tend not to choose to repeatedly recall material while they are studying.

Students' reports about the strategies they use in real world learning scenarios also indicate that the use of retrieval practice is rare. In one survey of college students by Karpicke, Butler, and Roediger (2009), 84% of students indicated that they repeatedly read as a study strategy, while only 11% indicated that they practiced actively recalling while they studied (see too Kornell & Bjork, 2007). In another survey, Wissman, Rawson, and Pyc (2012) asked students to imagine they were studying a stack of flashcards and indicate how they would decide to stop studying a given flashcard. Only 26% of students said they would practice an item until they could recall it multiple times (consistent with Karpicke's, 2009, experimental findings), while about 40% of students said they would continue until they could recall items only once before stopping. Thus, students do not frequently use retrieval practice as a study strategy, and when they do, they do not tend to practice repeated retrieval (additional retrieval beyond recalling items once).

Students Have Difficulty Monitoring the Accuracy of Their Own Responses

If students attempt retrieval while learning on their own, they must evaluate whether what they recall is correct, and this too is a source of difficulty for students (Dunlosky, Hartwig, Rawson, & Lipko, 2011; Dunlosky & Rawson, 2012; Rawson & Dunlosky, 2007). Wissman et al. (2012) found that if students reported that they try to recall during learning, 27% said they would not compare their own responses to objective answers and would instead rely on subjective feelings (e.g., the ease with which material could be recalled). Even under circumstances in which students are required to compare their own responses to objective answers, they often believe they have recalled material correctly when in fact they have not. In a stunning demonstration of this, Rawson and Dunlosky (2007) had students attempt to recall definitions and then score their own responses. The students were shown their responses along with the correct answers and were told to award full, partial, or no credit to their responses. Students assigned full or partial credit to responses that were completely incorrect 43% of the time. It is clear that overconfidence in the assessment of objectively incorrect responses would be detrimental in scenarios where students must monitor and regulate their own learning.

The research reviewed above points to several aspects of self-regulated learning that might conspire against students if they practice retrieval on their own. First, retrieval practice tends to yield lower judgments of learning relative to less effective activities like repeated rereading. Second, when students regulate their own study behaviors, they often do not practice retrieval at all, and if they do, they tend to recall items only once rather than practicing repeated retrieval. Third, students have great difficulty evaluating the accuracy of their own responses, often thinking they are correct when they are partially or sometimes completely incorrect. For these reasons, we carried out the following experiments on a new computer-based program that guides students to practice retrieval.

Introduction to the Experiments

The present article reports on our initial attempts to develop a computer program that guides students through retrieval practice of complex educational materials. This program is modeled after research using relatively simple materials (like foreign language vocabulary words) in which retrieval practice is controlled based on the correctness of student recall (e.g., Karpicke, 2009; Karpicke & Smith, 2012). In order to guide retrieval practice of more complex materials and student responses, we developed an automated scoring procedure called QuickScore, described in detail below. Computer-based scoring with QuickScore provides an objective measure of performance to guide learning, rather than relying on student's subjective and often inaccurate opinions of their performance. In Experiments 1A and 1B, we examined the effects of repeated retrieval and repeated studying of complex materials with QuickScore guiding the study and recall of particular items. In Experiment 2, we compared the scoring performance of QuickScore to students' self-scoring of their own responses during learning. We also examined whether QuickScore can be used to improve student self-scoring performance by highlighting missing parts of a student's answer during self-scoring.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Experiments 1A and 1B

In Experiments 1A and 1B, students learned anatomy materials in one of two learning conditions: repeated study or repeated retrieval. The materials were muscle attributes, such as the function, innervation, or location of a muscle group. In both conditions, students repeatedly alternated between study and recall periods (Karpicke, 2009; Karpicke & Bauernschmidt, 2011). In study periods, students studied retrieval cues and the muscle attributes (see Figure 1). In recall periods, students were given retrieval cues and recalled the attributes by typing into a text box (see Figure 1). After an attribute was correctly recalled, it was assigned either two additional recall trials (repeated retrieval) or two additional study trials (repeated study; Karpicke, 2009; Karpicke & Smith, 2012). Importantly, QuickScore evaluated the students' responses and determined when an attribute had been correctly recalled. Learning continued until all attributes had been dropped from the list. Thus, students in both conditions recalled all of the attributes at least one time and were equally exposed to the materials—the only difference was the number of retrieval opportunities provided. Students took a final test after 2 days in which they recalled the attributes (same format as initial recall periods). Based on previous research that used similar procedures but simpler materials (Karpicke & Bauernschmidt, 2011; Karpicke & Smith, 2012), we expected repeated retrieval to enhance long-term retention relative to repeated studying.

Method

Subjects and design. In total, 68 Purdue University undergraduate students participated in Experiments 1A and 1B. The students were enrolled in an introductory psychology course and participated in exchange for course credit. The age of students ranged from 18 to 42 years, although only two students were over the age of 23. The median age was 19.5 and 19 in Experiments 1A and 1B, respectively. Twenty-eight students participated in Experiment 1A, with 14 assigned to the repeated retrieval condition and 14 assigned to the repeated study condition. Forty students participated in Experiment 1B, with 20 assigned to the repeated retrieval condition and 20 assigned to the repeated study condition.

Materials. Two lists of nine muscle attributes were used (18 attributes total). The nine attributes in each list were the function, innervation, or location of three muscle groups (List 1: *deltoid muscle, triceps muscle,* and *digital flexor muscles*; List 2: *gluteus maximus muscle, quadriceps muscle,* and *gastrocnemius muscle*). Retrieval cues for each attribute were always the muscle group name plus "function," "innervation," or "location" (e.g., *deltoid muscle–function*; see the Appendix). Only List 1 was used in Experiment 1A, while both lists were used in Experiment 1B. Students in Experiment 1B always learned List 1 before List 2.

Procedure. Students were tested in small groups of up to four at a time. During an initial learning phase, students learned one list (Experiment 1A) or two lists (Experiment 1B) of muscle attributes by alternating between study and recall periods. During study periods, each attribute was presented on the screen one at a time for study. The materials were arranged on the screen so that the retrieval cues (e.g., *deltoid muscle–function*) were centered at the top of the screen with the attribute directly beneath it. When the student had finished studying, they clicked a button labeled "next" to advance to the next trial.

During recall periods, the retrieval cues were presented centered at the top of the screen with a text input box directly below them. Students were instructed to recall as much of the attribute as possible by typing into the response box. When the student was finished recalling, they pressed the return key or clicked a "next" button to submit their answer and advance to the next attribute.

The presentation order of attributes in the study and recall periods was blocked by muscle name, and the muscles were always presented in the same order (List 1: *deltoid muscle*, *triceps muscle*, then *digit flexor muscles*; List 2: *gluteus maximus muscle*, *quadriceps muscle*, then *gastrocnemius muscle*). However, the attributes for each muscle were presented in a random order.

The critical manipulation took place when an attribute was correctly recalled for the first time. In the repeated retrieval condition, once an attribute was correctly recalled it was assigned to be recalled two more times in the next two recall periods, but dropped from further study periods. In the repeated study condition, once an attribute was correctly recalled it was assigned to be studied two more times in the next two study periods but dropped

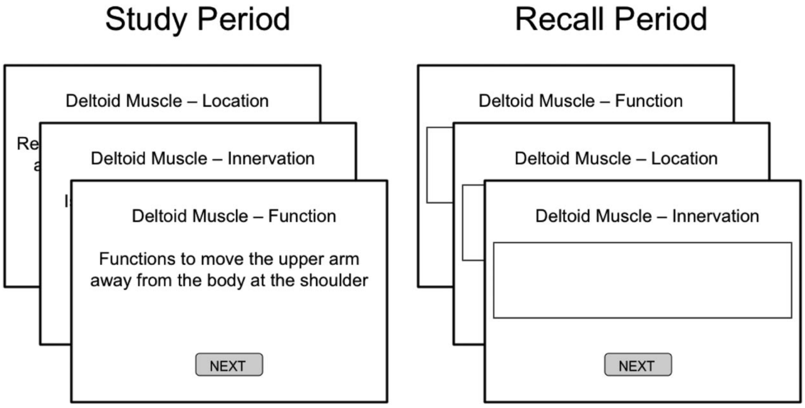

Study Period
Deltoid Muscle – Location
Deltoid Muscle – Innervation
Deltoid Muscle – Function
Functions to move the upper arm away from the body at the shoulder
NEXT
Recall Period
Deltoid Muscle – Function
Deltoid Muscle – Location
Deltoid Muscle – Innervation
NEXT

Figure 1. Example of the study periods and recall periods from Experiments 1A and 1B. During study trials, students studied the attributes one at a time. During recall trials, students attempted to retrieve the attributes. The order of muscle groups was held constant, but the order of attributes (i.e., location, innervation, function) was randomized.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

from further recall periods. With this manipulation, the total number of exposures to the attributes was matched, but the number of retrieval opportunities varied between conditions. Note that students in both conditions were required to retrieve each attribute at least one time.

After 2 days, students returned for a final test of the attributes. The final test was identical to one recall period during the first session. Students in Experiment 1B were tested on attributes from List 1, then on attributes from List 2.

Automated scoring. QuickScore evaluated each response by comparing it with the target attribute that was cued for recall. QuickScore's process for evaluating responses was as follows. First, any spelling errors in students' responses were corrected using contextualized spell correction (Leacock & Chodorow, 2003). For each incorrectly spelled word, a spell checker generated a list of plausible alternatives based on the characters used. If any of the suggestions were keywords in any of the target attributes, then the misspelled word was changed to that suggestion. Second, QuickScore determined the keywords of the attribute by using a list of *stop words*. Stop words are a list of function words that are typically noninformative (e.g., *and*, *the*, *to*). If a word in the attribute was not in the stop word list, then QuickScore considered it a keyword. Third, all words from the students' response and the keywords of the attribute were reduced to their stem using the Porter (1980) word-stemming algorithm (e.g., *muscle* and *muscles* were reduced to *muscl*). In theory, words with the same stem are likely to have similar meanings, and so this practice should improve scoring accuracy by emphasizing meaning over exact string matching. Finally, QuickScore tallied the number of keywords present in the response, and a score for the response was computed as the proportion of attribute keywords present in the response. If the response received a score of .75 or higher, then it was considered correct.

Experiment 1B featured an updated version of QuickScore that used WordNet (Fellbaum, 1998; Miller, 1995), a database of the English lexicon, to look up synonyms of the keywords prior to stemming. Synonyms of a keyword found in a response were counted as if the actual keyword was found.

Results

All results were significant at the .05 level unless otherwise stated.

Independent scoring. All responses from the learning phase and final test were scored by two independent human raters. Each response was scored as correct (1 point), partially correct (.5 points) or incorrect (0 points). The two raters agreed on 98% of their scores in Experiment 1A and 96% of scores in Experiment 1B. For responses on which the two raters did not agree, a third rater (Experiment 1A) or the first author (Experiment 1B) cast the deciding vote.

Learning phase performance. Figure 2 shows the cumulative learning curves from the initial learning phases in both experiments. The cumulative learning curve represents the proportion of attributes that had been correctly recalled at least one time on a given recall period (Karpicke, 2009; Karpicke & Roediger, 2007, 2008). No differences between the conditions were expected here, because assignment to condition occurred after an attribute was correctly recalled (e.g., Karpicke & Smith, 2012). As seen in Figure 2, both conditions learned the attributes at approximately the same rate. Both experiments were analyzed using a 2 (learning condition: repeated study or repeated retrieval) × 4 (recall period: 1–4) mixed analysis of variance (ANOVA). Only recall periods 1–4 were included because performance approached ceiling after period 4. There was a main effect of recall period in both experiments [1A: $F(3, 78) = 171.23$, $\eta_p^2 = .87$; 1B: $F(3, 114) = 333.46$, $\eta_p^2 = .90$], indicating that recall improved across recall periods. However, there was no main effect of learning condition [1A: $F(1, 26) < 1$; 1B: $F(1, 38) = 1.56$], nor a Learning Condition × Recall Period interaction [1A: $F(3, 78) < 1$; 1B: $F(3, 130) = 2.08$] in either experiment, indicating that recall performance across recall periods was the same for both the repeated study and repeated retrieval conditions.

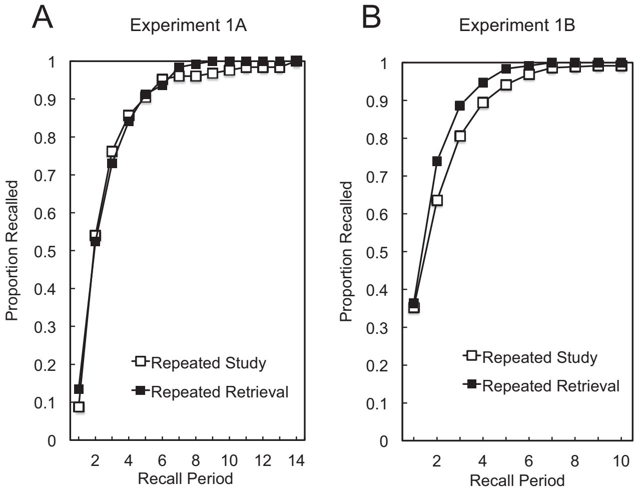


Figure 2. Cumulative recall of muscle attributes across recall periods during initial learning in Experiment 1A (A) and Experiment 1B (B). Students in both conditions learned the attributes at approximately the same rate and to the same criterion during the initial learning phase.

Final recall. Figure 3 shows the proportion of attributes recalled on the final test in Experiments 1A and 1B, respectively. In Experiment 1A, students in the repeated retrieval conditions recalled more attributes than students in the repeated study condition (.71 vs. .54), $t(26) = 2.02$, $d = 0.79$. The same result occurred in Experiment 1B: Students in the repeated retrieval conditions recalled more attributes than students in the repeated study condition (.70 vs. .58), $t(38) = 2.72$, $d = 0.87$. Thus, both experiments help establish that guided retrieval practice with automated scoring enhances long-term retention of relatively complex materials.

QuickScore performance. We examined QuickScore's performance during retrieval practice by comparing its scores to those of the independent raters. Because QuickScore was only used during the initial learning phase, data from the final test were not included in this analysis. Also, because QuickScore only scored items as correct or incorrect, the independent rater scores were transformed so that both partially correct and incorrect scores counted as incorrect. The relevant data are shown on Table 1. QuickScore agreed with the independent raters on 83% of trials in both Experiments 1A and 1B. We also computed kappa (κ) values, a measure of agreement accounting for chance, where a κ of 0 indicates chance agreement, and κ of 1 indicates perfect agreement. Kappa was .65 in Experiment 1A and .70 in Experiment 1B.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

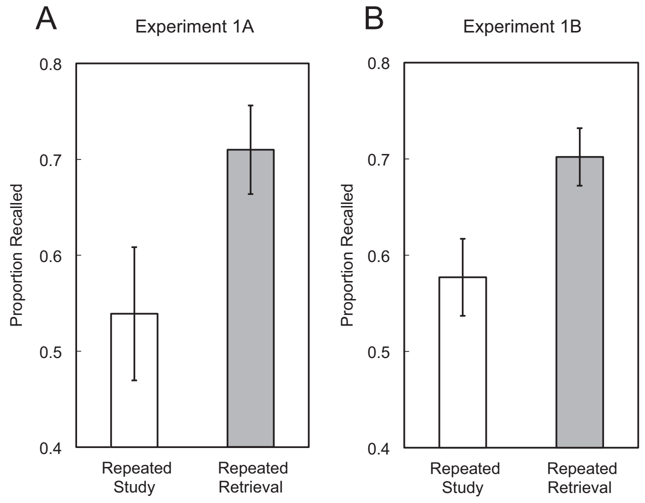


Figure 3. Proportion recalled following a 2-day delay in Experiment 1A (A) and Experiment 1B (B). Error bars represent standard error of the mean. Students in the repeated retrieval condition recalled more than students in the repeated study condition.

According to Fleiss and Paik (2003), κ values greater than .75 represent excellent agreement beyond chance, and κ values between .40 and .75 represent fair to good agreement beyond chance. Thus, agreement between QuickScore and the independent raters can be classified as good.

We also analyzed the types of errors that QuickScore made by examining false negatives and false positives, where false negatives referred to items scored as correct by independent raters but incorrect by QuickScore, and false positives referred to items scored as incorrect by independent raters but correct by QuickScore. These data are also shown on Table 1. QuickScore's false negative rate was .18 in Experiment 1A and .14 in Experiment 1B, and its false positive rate was .17 in Experiment 1A, and .17 in Experiment 1B. We inspected each of these errors to determine their common causes. For false negatives, we identified three types of errors: *synonymous errors, criterion errors, and spelling errors*. Synonymous errors were the most frequent (83%), and occurred when the response used phrasing that expressed the attribute completely, but using words that were not among the keywords used by QuickScore. Criterion errors were less frequent (12%), and occurred when the response successfully expressed the attribute using fewer keywords than required by QuickScore to be correct. Spelling errors were rare (5%), and occurred when misspelled words were missed by QuickScore's spell checker. For false positives, we identified two types of errors: *inaccurate errors* and *incomplete errors*. Inaccurate errors were the most frequent (65%), and occurred when the response used enough keywords for QuickScore to count as correct, but not in a way that was factually accurate. For example, a response of "bends the lower arm at the elbow" to the cue *triceps muscle–function* contains enough keywords of the target "extends the lower arm at the elbow" to count as correct but is factually inaccurate. Incomplete errors were less frequent (35%) and occurred when the response was factually accurate but incomplete. For example, a response of "extends the lower arm" to the cue *triceps muscle–function* does not mention to motion occurring at the elbow. It is important to note that the independent raters scored the majority of false positives as partially correct (76% and 85% in Experiments 1A and 1B, respectively), so QuickScore rarely gave full credit to completely incorrect responses.

Table 1
Agreement Between QuickScore and Independent Raters, Proportion of False Positives, and Proportion of False Negatives in Experiments 1A and 1B

Experiment	Agreement	False positives	False negatives
Experiment 1A	.83 (1,011)	.16 (570)	.18 (441)
Experiment 1B	.83 (2,775)	.20 (1,420)	.14 (1,355)

Note. "False positives" refers to the proportion of responses scored correct by QuickScore but incorrect by independent raters. "False negatives" refers to the proportion of responses scored incorrect by QuickScore but correct by independent raters. The number of responses contributing to the respective proportions is in parentheses.

The relationship between QuickScore and learning. Finally, we conducted a post hoc item analysis that examined the influence of QuickScore on final test performance. When QuickScore accurately evaluated a response, the attribute was appropriately assigned at the first correct recall and was presented for the predetermined number of recall opportunities (*on time* assignment). When QuickScore made a false positive, the attribute was assigned before the first correct recall and therefore received fewer recall opportunities (*early* assignment). When QuickScore made a false negative, the attribute was assigned after the first correct recall and therefore received additional recall opportunities (*late* assignment). To examine the influence of QuickScore on learning, we computed the proportion of attributes recalled at final test as a function of assignment during initial learning (early, on time, or late). The results are shown on Table 2.

There are two critical findings shown on Table 2. The first finding is that late assignment, caused by false negatives from QuickScore, resulted in the highest levels of recall. The second finding is that early assignment, due to false positives from QuickScore, resulted in the lowest levels of recall. These findings are readily explained by how early or late assignment altered the number of times attributes were retrieved during initial learning. Late assignment meant that an attribute could be recalled more

Table 2
Proportion of Muscle Attributes Recalled on Final Test as a Function of Assignment During Initial Learning

Variable	Early (False positives)	On time	Late (False negatives)
Experiment 1A			
Repeated Study	.28 (29)	.54 (70)	.83 (27)
Repeated Retrieval	.37 (23)	.77 (80)	.85 (23)
Experiment 1B			
Repeated Study	.53 (76)	.56 (242)	.79 (42)
Repeated Retrieval	.51 (69)	.72 (239)	.88 (52)

Note. Early assignment to a condition was caused by false positives and led to less practice. Late assignment was caused by false negatives and led to at least one extra trial of practice. Total number of attributes assigned early, on time, or late is in parentheses.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

than one time before it was assigned, which resulted in the enhanced long-term performance in both repeated study and repeated retrieval conditions. Conversely, early assignment meant that an attribute would never be correctly recalled in either condition. Note that if total study time was predictive of learning, then early assignment should favor the repeated study condition over repeated retrieval, as attributes assigned early in the repeated study condition would at least receive two extra restudy trials, while attributes in the repeated retrieval condition would continue to be incorrectly recalled without corrective feedback. This is not observed in Table 2. Of course, this type of post hoc analysis is subject to item selection artifacts, but nevertheless demonstrates how scoring can affect learning by altering the amount of retrieval during practice. False positives produced a negative effect on learning, while false negatives actually improved learning.

Discussion

In two experiments, we used an automated short-answer scoring algorithm, QuickScore, to replicate previous findings of retrieval enhanced learning with complex anatomy materials (Karpicke & Roediger, 2007, 2008). Students who repeatedly retrieved muscle attributes recalled more of the attributes after a 2-day delay than students who repeatedly studied. Moreover, the repeated retrieval manipulation produced medium to large effect sizes (d = 0.79 and 0.87 for Experiments 1A and 1B, respectively). Note that students in both conditions were brought up to the same criterion level of performance before the manipulation began—each student correctly recalled the attributes at least one time. The difference was whether the students continued to study or recall the attributes after they were correctly recalled. Repeated retrieval was the critical factor in enhancing performance on final test. Importantly, the retrieval manipulation was successfully implemented using QuickScore. QuickScore performed quite well, and achieved over 80% agreement with independent raters in both experiments. QuickScore's errors led to differential effects on learning. False negatives caused some attributes to be assigned late, which resulted in extra retrieval practice and better final test performance. Conversely, false positives caused some attributes to be assigned early, which resulted in less retrieval practice and worse final test performance. Again, repeated retrieval was the critical factor in enhancing performance on the final test.

Experiment 2

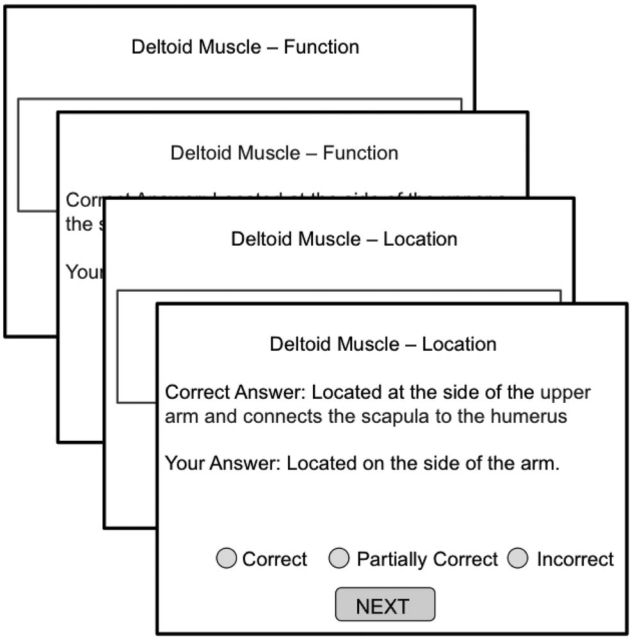
One of the primary motivations for developing QuickScore was to provide students with objective scoring during retrieval practice. Previous research had shown that when students self-score their own responses, they are often overconfident and make scoring errors (Dunlosky & Rawson, 2012; Dunlosky et al., 2011). However, it is unknown whether QuickScore provides better scoring than students themselves. The primary goal of Experiment 2 was to directly compare student self-scoring to QuickScore.

Our secondary goal was to examine whether QuickScore could be used as a way to improve student self-scoring. Dunlosky et al. (2011) proposed that self-scoring is challenging for students because it taxes their working memory abilities by requiring them to actively maintain which elements of a response have or have not been correctly recalled. In the present experiment, we used QuickScore to highlight keywords that were missing from the students' responses, as the students were self-scoring. We reasoned that highlighting the terms that were missing from a response should reduce the working memory demands of self-scoring by externalizing the information that students had to track. Thus, we predicted that highlighting missing key-terms with QuickScore should improve self-scoring performance.

The general procedure used in Experiment 2 was modified from the one used in Experiments 1A and 1B. Students studied muscle attributes in an initial study period and then attempted to recall the attributes in a recall period. Unlike previous experiments, each recall trial was followed immediately by a self-score trial in which the students self-scored their response by comparing it to the correct answer (see Figure 4). After QuickScore determined that an attribute was correctly recalled it was dropped from the list and not presented for any additional recall trials. Recall periods were repeated until all attributes were dropped from the list. Students learned two lists of attributes in this manner. For one of the lists, students received highlighted feedback from QuickScore during self-scoring. On self-score trials for this list, the target attribute had keywords that were missing from the student's response printed in red. Students were instructed to use the highlighted words to help them score their response. Note that unlike Experiments 1A and 1B, we did not include a delayed test or retrieval manipulation, as the focus of Experiment 2 was on self-scoring.

Method

Subjects, materials, and design. Thirty-four Purdue University undergraduate students from an introductory psychology course participated in exchange for course credit. None of the students had participated in Experiments 1A or 1B. The age of



Deltoid Muscle – Function

Deltoid Muscle – Function

Cor

the s

You

Deltoid Muscle – Location

Deltoid Muscle – Location

Correct Answer: Located at the side of the upper arm and connects the scapula to the humerus

Your Answer: Located on the side of the arm.

Figure 4. Example of recall period with self-score trials from Experiment 2. Immediately after each recall trial, students were asked to restudy the correct answer and self-score their response.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

students ranged from 18 to 24, and the median age was 19 years. Students learned two lists of muscle attributes, which were the same materials used in Experiment 1B. Whether the correct answers were highlighted or not highlighted during self-score trials was manipulated within subjects, with one list assigned to the highlighting condition and the other assigned to the no highlighting condition. The order of highlighting conditions was counterbalanced across subjects.

Procedure. Students were tested in small groups of up to four at a time. Students learned each of the two lists separately, as in Experiment 1B. For each list, the students studied the attributes in an initial study period study. After the initial study period, the students attempted to retrieve the attributes in a recall period (see Figure 4). Immediately after each recall attempt, the student was given a self-score trial of the attribute. During the self-score trial, the retrieval cue (e.g., *deltoid muscle–function*) was centered at the top of the screen, the target attribute was directly beneath the retrieval cue, and the response was directly beneath the target attribute. The target attribute and response were labeled so that the student could tell the two apart (as “Correct Answer” and “Your Answer,” respectively). If the list was scheduled to receive highlighted feedback, QuickScore highlighted keywords that were missing from the student’s response. Highlighted words appeared in red print. At the bottom of the screen were three radio buttons labeled “correct,” “partially correct,” and “incorrect” and a button labeled “next.” Students were instructed to restudy the correct answer and to score their own response using the radio buttons and click “next” when they were finished. After clicking next they were moved on to a recall trial of the next attribute.

After the students had attempted recall of all attributes in the list, the recall period was repeated. All responses were scored on-the-fly by QuickScore, using the same version of QuickScore as Experiment 1B. If QuickScore determined that an attribute was correctly recalled, it was dropped from the list at the end of the recall period. Recall periods were repeated until all attributes had been dropped from the list.

Results

Independent scoring. Two independent raters scored all responses using the same method as the previous experiments. The two raters agreed on 93% of their scores, and a third rater resolved any scoring discrepancies between the two raters by casting the deciding vote.

Learning phase performance. Figure 5 shows the cumulative learning curves from the initial learning phase. As seen in Figure 5, students learned the attributes at approximately the same rate in both highlight conditions. These results were analyzed using a 2 (highlight: highlight, no highlight) × 4 (recall period: 1–4) repeated-measures ANOVA. Only recall periods 1–4 were included because recall was at ceiling after the fourth period. The main effect of recall period was significant, $F(3, 99) = 202.52$, $\eta_p^2 = .86$, indicating that recall improved across recall periods. Neither the main effect of feedback, $F(1, 33) < 1$, nor the Learning Condition × Recall Period interaction, $F(3, 99) < 1$, was significant, indicating that highlighting missed words during restudy produced no benefit to learning of the attributes.

Student scoring and QuickScore performance. We examined scoring performance for both student self-scoring and QuickScore by calculating the same measures as previous experiments. Student self-scores were converted to the same scale as QuickScore by treating self-scores of “incorrect” and “partially correct” as “incorrect.” For direct comparisons between proportion scores, the typical ANOVA approach is not appropriate, as each subject produced a different number of correct and incorrect responses and therefore received a different number of self-scoring trials. Instead, we used nonparametric chi-squared tests that were adjusted to account for the different number of trials among subjects and to account for within subject correlation among observations (e.g., Fleiss & Paik, 2003; Rao & Scott, 1992).

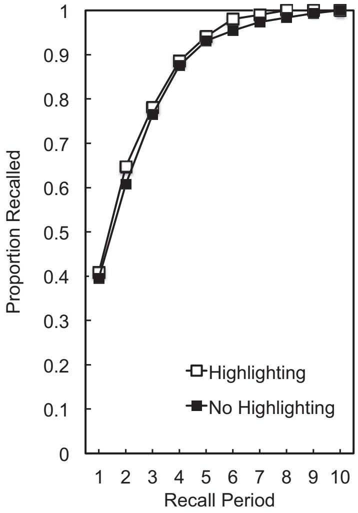


Figure 5. Cumulative recall of muscle attributes across recall periods during in Experiment 2. Students learned attributes that were highlighted during self-scoring at the same rate as attributes that were not highlighted.

First, we examined overall scoring agreement with the independent raters, for both QuickScore and the students. These data are shown in Table 3. QuickScore agreed with the independent raters on 78% of trials ($\kappa = .56$), while the students (collapsed across highlight condition) agreed with the independent raters on 72% of trials ($\kappa = .46$). This difference in agreement between QuickScore and the students was significant, $\chi^2(1) = 9.47$. Thus, in terms of overall scoring performance, QuickScore performed better than students self-scoring their own responses. We also examined whether providing highlighted feedback improved self-scoring performance. These data are shown in Table 3. Agreement with the independent raters was slightly better when students were given highlighted feedback (74%, $\kappa = .50$) than when no highlighted feedback was given (69%, $\kappa = .42$). However, this advantage was not significant, $\chi^2(1) = 1.26$, $p = .26$. Thus, highlighting during self-scoring did not provide an improvement in the overall agreement between students and the independent raters.

Next, we examined the types of scoring errors that QuickScore and students made. We calculated a false negative rate and false positive rate for QuickScore and students, as in Experiment 1. The results are shown in Table 3. Table 3 shows that errors made during learning were much different between students and QuickScore. QuickScore was more likely to commit false negatives than students (.26 vs. .08), $\chi^2(1) = 42.09$. However, QuickScore was also less likely to commit false positives than students (.18 vs. .42), $\chi^2(1) = 64.63$. Thus, while both QuickScore and students made

Table 3
Agreement With Independent Raters, Proportion of False Positives, and Proportion of False Negatives for Student Self-Scoring and QuickScore in Experiment 2

Variable	Agreement	False positives	False negatives
Student Self-Scoring			
Highlighting	.74 (724)	.36 (433)	.11 (291)
No Highlighting	.69 (771)	.48 (458)	.05 (313)
Overall	.72 (1,495)	.42 (891)	.08 (604)
QuickScore	.78 (1,495)	.18 (891)	.26 (604)

Note. Highlighting indicates trials on which students were provided with highlighted feedback during self-scoring. The total number of responses contributing to the respective proportions is in parentheses.

errors during scoring, they each committed false negatives and false positives at different rates. We also examined whether highlighted feedback during self-scoring improved scoring in terms of the types of errors made. As seen in Table 3, the students' false negative rate was slightly higher on trials with highlighted feedback than on trials without highlighted feedback, $\chi^2(1) = 4.42$. The students' false positive rate was numerically lower on trials with highlighted feedback than on trials without highlighted feedback; however, this difference did not reach significance, $\chi^2(1) = 3.16$, $p = .075$.

Discussion

The primary goal of this experiment was to compare the scoring performance of QuickScore to that of students scoring their own responses. In terms of overall agreement with the independent raters, QuickScore was more accurate than student self-scoring. In terms of the type of scoring errors made, QuickScore was more likely to commit false negatives than the students, but less likely to commit false positives than students. The high false positive rate and low false negative rate observed in student self-scoring data is consistent with previous research (e.g., Rawson & Dunlosky, 2007) and indicates a general bias to give credit to one's responses when self-scoring.

The secondary goal of this experiment was to determine whether QuickScore could be used to improve student self-scoring by providing highlighted feedback. Providing students with highlighted feedback during self-scoring produced no improvement in the overall agreement with the independent raters, failing to support the idea that reducing working memory demands would improve self-scoring. However, highlighted feedback did alter the types of errors committed by students—students were less likely to false positive and more likely to false negative when given highlighted feedback. We hasten to add that these observed differences were small or did not reach statistical significance. Nevertheless, the pattern of errors suggests that highlighted feedback made students more conservative in their scoring, perhaps by reducing the bias to give credit to responses.

General Discussion

Retrieval practice is a powerful way to enhance student learning. However, students are unaware of the direct benefits of retrieval (e.g., Roediger & Karpicke, 2006) and typically view retrieval only as a way to assess their knowledge. As a result, students do not choose to practice retrieval as much as they should and tend to discontinue retrieval practice after they can recall something one time (e.g., Karpicke, 2009). Moreover, students have difficulty evaluating the correctness of their retrieval attempts, and often think they have correctly recalled something when they have not (Dunlosky et al., 2011; Rawson & Dunlosky, 2007). These findings suggest that students would benefit from guidance during retrieval practice. In the present study, we examined the effectiveness of a computer program that guides retrieval practice of anatomy materials using an automated scoring algorithm, QuickScore. During guided retrieval practice, QuickScore evaluates the correctness of students' retrieval attempts and uses this information to make decisions about when students should discontinue studying or retrieving the materials. We review some of the main findings from the study, and discuss some of their practical and theoretical implications.

Guided Retrieval With Automated Scoring

Repeated retrieval enhanced long-term retention, and QuickScore was effective in guiding retrieval practice. In Experiments 1A and 1B, we tested whether QuickScore was capable of guiding retrieval practice by using it to replicate previous retrieval practice research, using complex anatomy materials instead of word pairs. In these experiments, students practiced retrieval by alternating between study and retrieval phases of muscle attributes. When QuickScore determined that an attribute was correctly recalled, it was assigned to either two additional recall trials (repeated retrieval) or two additional study trials (repeated study; Karpicke, 2009; Karpicke & Smith, 2012). Students in the repeated retrieval conditions recalled more attributes after a delay than students in the repeated study conditions (Karpicke & Roediger, 2008). These results add to a growing literature showing the importance of retrieval for learning and demonstrate that QuickScore is capable of guiding retrieval practice.

QuickScore showed good correspondence with objective, independent raters. Overall agreement with the independent raters ranged between 78% and 83% across all three experiments. This is a typical level of performance for this class of short-answer scoring algorithms (Leacock & Chodorow, 2003). The types of scoring errors made by QuickScore can be split into two classes, false negatives and false positives. QuickScore committed false negatives and false positives with an equal frequency across the experiments. False negatives were most often the result of students using words and phrases that expressed the attribute without actually using the keywords. False positives were most often the result of students using enough keywords to be correct, but in a way that was either factually inaccurate or incomplete. Such problems are common among keyword based scoring algorithms that do not account for the overall meaning implied by the arrangement of words. Other scoring algorithms have implemented advanced natural language processing techniques to successfully deal with issues related to word order (e.g., Leacock & Chodorow, 2003). We are currently exploring other ways of improving QuickScore's accuracy, such as using part-of-speech taggers to exclude uninformative adverbs (e.g., "mainly" and "mostly"), weighting keywords according to normative word frequency, and adjusting the number of keywords required for each attribute.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Scoring accuracy during retrieval practice was important, because attributes were assigned to conditions according to whether they were recalled correctly. Our results suggest that not all scoring errors are necessarily bad in a retrieval practice program and can sometimes actually improve learning. QuickScore's false negatives led to late assignment of some attributes, which give them extra retrieval practice. These attributes had the highest levels of recall at final test. In contrast, QuickScore's false positives led to early assignment of some attributes, which ended practice before ever being correctly recalled. These attributes had the lowest levels of recall at final test. Our view is that a conservative approach to scoring, which may result in more false negatives but fewer false positives, is desirable because it will lead to more practice, whereas false positives would lead to early assignment of attributes to conditions, and practice would end before attributes are fully learned. This general principle can help guide researchers and practitioners in the future development of retrieval practice applications. When choosing or developing scoring algorithms, or evaluating student self-scoring performance, false positives rates should be regarded as more important than overall accuracy.

It is important to note that QuickScore is only one of many automated scoring systems that have been developed over the last few decades (for a review, see Pérez-Marín, Pascual-Nieto, & Rodríguez, 2009). These systems use a variety of sophisticated methods for scoring, such as natural language processing (e.g., Leacock & Chodorow, 2003; Sukkarieh, Pulman, & Raikes, 2004), latent semantic analysis (e.g., Foltz, Laham, & Landauer, 1999), n-gram co-occurrence (Noorbehbahani & Kardan, 2011), and machine learning (e.g., Mohler, Benescu, & Mihalcea, 2011; Nehm, Ha, Mayfield, 2012). The intended purpose of most automated scoring systems is formative assessment, but any of these systems could also be used in a retrieval-based learning program, like QuickScore. Indeed, latent semantic analysis has been used with great success in automated tutoring programs (e.g., Graesser, Penumatsa, Ventura, Cai, & Hu, 2007; McNamara, Boothum, Levinstein, & Millis, 2007), as well as programs designed to help summarization and writing skills (Kintsch, Steinhard, Stahl, & the LSA Research Group, 2000; Wiemer-Hastings & Graesser, 2000). However, all automated scoring systems are tailored to particular types of responses (e.g., essays, summarizations, short answer questions, etc.), and their performance varies across materials and response types. Thus, determining which system is "best" will depend to a large extent on the types of responses under consideration, as well as the materials being used. QuickScore worked reasonably well in the present study, but we recommend that researchers working to apply automated assessment to retrieval practice consider the available options and use a system that works for their unique application.

Student Self-Scoring

In Experiment 2, we examined the scoring performance of both QuickScore and student self-scoring by comparing each to independent raters. Overall, QuickScore agreed with the independent raters more often than students. These results align with previous studies showing that students are inaccurate when self-scoring their own responses (e.g., Dunlosky et al., 2011; Dunlosky & Rawson, 2012; Rawson & Dunlosky, 2007). Interestingly, QuickScore and students produced very different false negative rates and false positive rates. QuickScore had a higher false negative rate than students but also had a lower false positive than students. Thus, while QuickScore was relatively conservative, the students were more liberal. In our previous analysis of QuickScore, we concluded that conservative scoring was more desirable during retrieval practice. It follows that using QuickScore to guide retrieval practice would result in better learning than using students' own self-scoring.

Why do students make errors when self-scoring? One idea is that errors occur when the cognitive demands of self-scoring exceeds the working memory capacity of the students (Dunlosky et al., 2011). We tested this idea in Experiment 2 by using QuickScore to highlight keywords that were missing in the students' response. Our rationale was that the highlighted terms would provide students with an external visual indicator of their performance, thereby reducing the working memory requirements of self-scoring and improving self-scoring performance. However, highlighting keywords did not improve the agreement between the students and the independent raters. Thus, the results do not support the idea that errors are caused by limited working memory.

While we did not find evidence in support of the limited working memory idea, the data from Experiment 2 do hint to an alternate explanation for self-scoring errors. First, students were much more likely to false positive than false negative when self-scoring. This pattern suggests that students are biased to award credit to responses, regardless of objective information in the response. If errors were due to an inability to keep track of and hold objective information in mind, then false negatives and false positives should occur at equal rates. Second, highlighted feedback actually *increased* the false negative rate, meaning that students were less likely to give credit to objectively correct responses. This suggests that highlighted feedback acted to make the students more conservative, perhaps by shifting the subjective criterion used during self-scoring. We can only speculate as to what metacognitive cues students use when they self-score and leave this as a question for future research.

Conclusion

Computer technology has become ubiquitous in educational environments, and the time is ripe to explore ways of implementing this technology in the context of established learning techniques. A variety of sophisticated scoring algorithms, like the one used in this experiment, have been developed to evaluate short-answer responses (e.g., Leacock & Chodorow, 2003; Mohler & Mihalcea, 2009; Sukkarieh et al., 2004). However, this technology tends to be used for testing and assessment purposes only. While assessment is an important to education, we see greater educational potential for using this technology to aid students within computerized learning environments. One successful example of this is AutoTutor, which uses latent semantic analysis in an immersive tutoring environment (Graesser et al., 2007). Our approach is novel in that we have used automated scoring to promote learning through active retrieval practice. We believe that combining the powerful learning produced by retrieval practice with sophisticated scoring algorithms could prove to be particularly potent way to enhance student learning.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876. doi:10.1002/acp.1391

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1118–1133. doi:10.1037/a0019902

Dunlosky, J., Hartwig, M., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology, 64,* 467–484. doi:10.1080/17470218.2010.502239

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

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fleiss, J. L., & Paik, L. B. (2003). *Statistical methods for rates and proportions*. New York, NY: Wiley. doi:10.1002/0471445428

Foltz, P., Laham, D., & Landauer, T. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2). Retrieved from http://imej.wfu.edu/index.asp

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 243–262). Mahwah, NJ: Erlbaum.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101,* 621–629. doi:10.1037/a0015183

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

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science, 21,* 157–163. doi:10.1177/0963721412443552

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1250–1257. doi:10.1037/a0023436

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331,* 772–775. doi:10.1126/science.1199327

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471–479. doi:10.1080/09658210802647009

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*. Advance online publication. doi:10.1007/s10648-012-9202-2

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. doi:10.1016/j.jml.2006.09.004

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968. doi:10.1126/science.1152408

Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition, 38,* 116–124. doi:10.3758/MC.38.1.116

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67,* 17–29. doi:10.1016/j.jml.2012.02.004

Kintsch, E., Steinhard, D., Stahl, G., & the LSA Research Group. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments, 8,* 87–109. doi:10.1076/1049-4820(200008)8:2;1-B;FT087

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224. doi:10.3758/BF03194055

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17,* 493–501. doi:10.1080/09658210902832915

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37,* 389–405. doi:10.1023/A:1025779619903

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., . . . Zhang, H. (2009). Clickers in the classroom: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology, 34,* 51–57. doi:10.1016/j.cedpsych.2008.04.002

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399–414. doi:10.1037/a0021782

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

McNamara, D. S., Boonthum, C., Levinstein, I., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 227–241). Mahwah, NJ: Erlbaum.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38,* 39–41. doi:10.1145/219717.219748

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 752–762). Portland, OR: Association for Computational Linguistics.

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short-answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL* (pp. 567–575). Athens, Greece: Association for Computational Linguistics.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education Technology, 21,* 183–196. doi:10.1007/s10956-011-9300-9

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers & Education, 56,* 337–345. doi:10.1016/j.compedu.2010.07.013

Pérez-Marín, D., Pascual-Nieto, I., & Rodríguez, P. (2009). Computer-assisted assessment of free-text answers. *The Knowledge Engineering Review, 24,* 353–374. doi:10.1017/S026988890999018X

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14,* 130–137. doi:10.1108/eb046814

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437–447. doi:10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi:10.1126/science.1191465

Rao, J. N. K., & Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics, 48,* 577–585. doi:10.2307/2532311

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

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283–302. doi:10.1037/a0023956

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17,* 382–395. doi:10.1037/a0026252

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. doi:10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2004). *Auto-Marking 2: An update on the UCLES-Oxford University research into using computational linguistics to score short, free text responses*. Philadelphia, PA: International Association of Educational Assessment.

Wiemer-Hastings, P., & Graesser, A. (2000). Select-a-kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments, 8,* 149–169. doi:10.1076/1049-4820(200008)8:2;1-B;FT149

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review, 18,* 1140–1147. doi:10.3758/s13423-011-0140-7

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix

Materials Used in Experiments

Retrieval cue	Attribute
	List 1
The Deltoid Muscle	
Function	• Functions mainly to move the upper arm away from the body at the shoulder
Innervation	• Is stimulated by the axillary nerve
Location	• Resides on the side of the upper arm at the shoulder and connects the scapula to the humerus
The Digital Flexor Muscles	
Function	• Function together to flex the fingers causing the formation of a fist
Innervation	• Is stimulated by the median and ulnar nerves
Location	• Reside on the front of the forearm when the palm is facing forward and connects the arm to the fingers
The Triceps Muscle	
Function	• Functions to extend the lower arm at the elbow
Innervation	• Is stimulated by the radial nerve
Location	• Resides on the back of the arm and connects the upper arm to a bone of the lower arm
	List 2
The Gluteus Maximus Muscle	
Function	• Functions to extend the thigh at the hip
Innervation	• Is stimulated by the inferior gluteal nerve
Location	• Resides on the back and side of the pelvis and connects the pelvis to the thigh
The Quadriceps Muscle	
Function	• Functions to extend the lower leg at the knee
Innervation	• Is stimulated by the femoral nerve
Location	• Resides on the front of the thigh and connects the upper leg to the kneecap and a bone of the lower leg
The Gastrocnemius Muscle	
Function	• Functions to extend the foot at the ankle
Innervation	• Is stimulated by the tibial nerve
Location	• Resides on the back of the lower leg and connects the leg to the heel

Received November 12, 2012
Revision received March 26, 2013
Accepted April 8, 2013 ■