Ψ Psychology Press
Taylor & Francis Group

A cognitive-science based programme to enhance study efficacy in a high and low risk setting

Janet Metcalfe

Department of Psychology, Columbia University, New York, NY, USA

Nate Kornell

Department of Psychology, University of California, Los Angeles, USA

Lisa K. Son

Barnard College, New York, NY, USA

In three experiments, learning performance in a 6- or 7-week cognitive-science based computer-study programme was compared to equal time spent self-studying on paper. The first two experiments were conducted with grade 6 and 7 children in a high risk educational setting, the third with Columbia University undergraduates. The principles the programme implemented included (1) deep, meaningful, elaborative, multimodal processing, (2) transfer-appropriate processing, (3) self-generation and multiple testing of responses, and (4) spaced practice. The programme was also designed to thwart metacognitive illusions that would otherwise lead to inappropriate study patterns. All three experiments showed a distinct advantage in final test performance for the cognitive-science based programme, but this advantage was particularly prominent in the children.

A basic foundation for school accomplishment is the availability of higher order cognitive and metacognitive competencies to realistically assess one's knowledge, to allocate and organise study time and effort optimally, to apply cognitive principles (such as deep, elaborative rehearsal, self-generation, testing, and spacing of practice, to name just a few) that effectively enhance

learning, and to resist the distractions that could undermine even the most sincerely endorsed intentions. In the population that we targeted in this research, a population of middle school children in an at-risk school setting, these competencies were strikingly limited. Their enhancement was the primary objective of this research. In addition, even among sophisticated learners, specific limitations exist in the use of certain metacognitive strategies. Children and adults often think they know things when they do not (see Metcalfe, 1998; Rawson & Dunlosky, 2007 this issue), and hence inappropriately terminate self-controlled study efforts, or otherwise study in a manner that fails to optimise learning (Bjork, 1994). We sought to devise a computer-assisted study programme, based on principles of cognitive science, that would allow such metacognitive illusions to be overcome.

The project reported here focused on memory enhancing principles derived from experimental studies in cognitive science. Most of these principles, because they have been studied primarily with a focus on understanding the mechanisms underlying memory and cognition, rather than on efforts to facilitate children's academic success, have been investigated in single-session experiments with scholastically irrelevant materials with college-student participants. Despite the apparent lack of connection of this cognitive science literature to real problems that children face in school, to the extent that the principles of cognitive science have some generality, we posited that by implementing them we should be able to enhance learning. The particular cognitive-science principles that we endeavoured to implement included (1) meaningful, elaborative, multimodal processing, (2) test-specific or transfer-appropriate processing, (3) self-generation and multiple testing of responses, and (4) spaced practice. In each of these areas, we focused on overcoming maladaptive metacognitive illusions that might otherwise mislead the student into studying inappropriately. We elaborate on each of these principles below, and review the most relevant research literature.

MEANINGFUL, ELABORATIVE MULTIMODAL PROCESSING

It is now well established that when people process material shallowly—at a perceptual, rote, or nonsemantically informed level—their memory is worse than if they process it deeply, meaningfully, or semantically (Cermak & Craik, 1979; Craik & Lockhart, 1972; Craik & Tulving, 1975), although the explanation for this levels-of-processing effect is still debated (Baddeley, 1978; Metcalfe, 1985, 1997; Nelson, 1977). Memory is critically dependent on people's having a schematic framework providing the deep meaning for the material without which memory performance deteriorates. For example, Bransford and Johnson (1973) showed that (1) when a schema was presented

before a story, subjects' comprehension ratings were high and their memory for the story was also high, (2) when the schema was not presented at all subjects' comprehension ratings were low and their memory was low, and (3) when the schema was presented *after* the story, there was a mismatch between the subjects' metacognitive judgements and their performance: Comprehension was rated as high, but memory was low. This apparent conflict between the individual's metacognition and performance may be especially problematic for children, who may feel that once they have understood a set of material, no further study is needed. Hence, in the computer controlled condition we devised a study programme that emphasised the meaningful deep structure of the materials, and provided elaborate rich examples. But to thwart the metacognitive illusion, we also tested the children and reexposed them to the to-be-learned material even in circumstances where they might spontaneously deem it unnecessary.

College students, in a free recall situation, rehearse by interweaving items (Metcalfe & Murdock, 1981; Murdock & Metcalfe, 1978; Rundus, 1971) into meaningful, coherent stories and images. In contrast, Korsakoff amnesics—whose memory performance is seriously deficient—rehearse repetitiously only the last item presented (Cermak, Naus, & Reale, 1976; Cermak & Reale, 1978; and see Metcalfe, 1997)—a strategy that has been shown, even with normal participants, to have little beneficial effect on memory (Bjork, 1970, 1988; Geiselman & Bjork, 1980). Children, who have little experience with studying, may be unaware of the need for integrative and meaningful rehearsal, and may feel that mere rote repetition is enough. Lack of insight into the need for elaborative rehearsal seems especially likely since it has been shown that feelings of knowing can be increased by mere priming (e.g., Reder, 1987; Schwartz & Metcalfe, 1992), and hence this aspect of effective encoding is critical. Moreover, people fail to realise that elaborative rehearsal is more effective than maintenance rehearsal (Shaughnessy, 1981). In the computer programme we devised, to overcome these metacognitive illusions, we mimicked meaningful, elaborative rehearsal by presenting materials in varied contexts, and by using multimodal presentation. We also presented items in several different contexts rather than just one.

TRANSFER APPROPRIATE PROCESSING, OR ENCODING SPECIFICITY

Tulving (1983; Tulving & Thomson, 1973; and more recently, see Roediger, Gallo, & Geraci, 2002) has emphasised that encoding is effective only to the extent that it overlaps with the operations required at the time of retrieval. This encoding specificity principle is inherent in most formal theories of human memory (e.g., Hintzman, 1987; Metcalfe, 1982, 1985, 1995) and has

considerable empirical support (e.g., Fisher & Craik, 1977; Hannon & Craik, 2001; Lockhart, Craik, & Jacoby, 1976). Metacognitive research, however, suggests that people are unlikely to take the test situation into account on their own. For example, it is well-established that people make judgements of learning based on heuristics such as how easy they find it to recall the to-be-learned information at the time the judgement is made (Benjamin, Bjork, & Schwartz, 1998; Dunlosky & Nelson, 1994; Dunlosky, Rawson, & McDonald, 2002; Kelley & Lindsay, 1993; Metcalfe, 2002; Weaver & Kelemen, 1997). But if the test conditions do not correspond to the conditions at the time of judgements, the judgements will be inaccurate, and people may, therefore, study inappropriately. For example, when Dunlosky and Nelson (1992) provided participants with targets as well as cues, judgements of learning were inaccurate. Similarly, Jacoby and Kelley (1987) asked participants for their ratings of objective difficulty on a task that was intrinsically very difficult (solving anagrams). Those subjects who were given unsolved anagrams were much more accurate (and thought the problems were more difficult) than people who were given the solutions. These results are important because children, while studying, may routinely look up answers, rather than trying to produce them. Such a strategy will induce illusory confidence. In summary, children may put themselves into the position of believing that they know information when in truth they do not, because they do not test themselves. The study programme implements a self-testing procedure that closely mimics the criterion test itself, and is thereby designed to overcome this illusion.

SELF-GENERATION OF RESPONSES

There is considerable support for the idea that learning is facilitated when people actively attempt to remember and generate responses themselves, rather than passively processing information spoon-fed to them by someone else. In experiments in which college students generated (as compared to read) free associates or rhymes, memory was enhanced (e.g., Schwartz & Metcalfe, 1992; Slamecka & Graf, 1978). The effect has been shown to occur with words (Jacoby, 1978), sentences (Graf, 1980), bigrams (Gardiner & Hampton, 1985), numbers (Gardiner & Rowley, 1984), and pictures (Peynircioglu, 1989), so long as the format of the test is the same as the format of study (Johns & Swanson, 1988; Nairne & Widner, 1987). In most studies, the items have been carefully selected such that in the generate condition, the participant always generated the correct answer. So, for example, in the generate condition, the cue might be Fruit: B_____, and the person had to say banana, whereas in the read condition the stimulus was Fruit: Banana, and the person had only to read the pair. Many studies

have shown effects favouring the generation condition. These effects have also been obtained when the answer was not so obvious, and required effortful retrieval from memory, which has been shown to lead to more learning than simply being presented with the answer (Bjork, 1975, 1988; Carrier & Pashler, 1992; Cull, 2000).

Thus far, though, there has been only one study investigating whether students opt to self-generate (Son, 2005). In that study, first graders made judgements of difficulty on cue–target vocabulary pairs. Then they were presented with only the cues and asked if they wanted to read or generate the targets. Results showed that the learners chose to self-generate, particularly when they judged the item to be easy. It remains unknown, however, whether the children would have opted to self-generate had both the cue and target been present when they made the choice. For example, when studying from a textbook, all of the information is present, and therefore, students may be unable to effectively self-generate, insofar as the presence of the answer thwarts efforts to generate it independently. By reading the answer, they have inadvertently put themselves into the "read" condition, which has been shown to be disadvantageous. We used a self-generation procedure in the current study programme.

SPACED PRACTICE

One of the most impressive manipulations, shown repeatedly to enhance learning, is spaced practice. There is now a large literature indicating that, for a wide range of materials, when items are presented repeatedly, they are remembered better if their presentations are spaced apart rather than massed (for a review see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). This effect may be of considerable pedagogical importance insofar as it has been shown by Bahrick and Hall (2005; and see Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005) to hold over extremely long periods of time, up to many years.

An "exception" to the spacing effect has been observed when the test occurs very soon after the last presentation of massed pairs (Glenberg, 1976, 1979). This reversal results in a compelling metacognitive illusion, however, and one that may be exceedingly difficult for the individual to overcome. Because of the immediate efficacy, people trained under massed practice are better pleased with the training procedure and give more favourable judgements about their learning than do those trained under spaced practice (Baddeley & Longman, 1978; Simon & Bjork, 2001; Zechmeister & Shaughnessy, 1980), even though eventual performance favours the latter group. For this reason, even adults, let alone children, do not intuitively understand the benefits of spaced practice, and spacing is unlikely to be used spontaneously; for example, recent data from first graders has shown

that children tend to choose massing strategies over spacing strategies (Son, 2005). Therefore, we implemented spaced practice in our learning programme. Two types of spacing were implemented, both of which have been shown to have positive effects: Spacing of items within a study session, and spacing of study sessions themselves.

METACOGNITION AND MOTIVATION

While many well-established findings within the scientific study of human memory show promise of contributing to the effectiveness of the study programme, people's metacognitions are often systematically distorted—mitigating *against* the spontaneous adoption of these effective study strategies. The past 10 years of intensive research has resulted in a growing understanding of the heuristic basis of the mechanisms underlying meta-cognitive judgements (Bjork, 1999; Dunlosky et al., 2002; Koriat, 1994, 1997; Koriat & Goldsmith, 1996; Metcalfe, 1993; Son & Metcalfe, 2005). Metacognitions do not appear to depend upon a person's privileged access to the true state of their future knowledge, but instead, on the information that the person has at hand at the moment the judgement is made. When that information is nondiagnostic or misleading, the judgements will be incorrect, and may lead to inappropriate study.

Deep understanding is both essential for memory and a goal of education. To extract, understand, appreciate, and articulate the core deep meaning of learning material, students need to be able to judge their learning in relation to an ideal state of deep understanding. Because surface knowledge and fluency can influence judgements of knowing, such a result is not automatic. For example, Metcalfe, Schwartz, and Joaquim (1993) showed that repeating a cue paired with the wrong target—a condition that hurt memory performance—resulted in inflated feeling-of-knowing judgements. Reder (1987) and Schwartz and Metcalfe (1992) found that priming parts of a question (e.g., ``precious stone'') increased subjects' feeling of knowing (e.g., on ``What precious stone turns yellow when heated?'') without changing the likelihood of the correct answer. Reder and Ritter (1992) found that exposure to parts of an arithmetic problem (e.g., $84 + 63$) inappropriately increased subjects' feeling of knowing to a different problem with surface overlap (e.g., 84×63). Oskamp (1962) provided psychiatrists and psychiatric residents with nondiagnostic information about a patient. The *irrelevant* information increased confidence without changing the accuracy of the diagnosis. The impact of surface knowledge on people's feelings-of-knowing is pernicious, because the inappropriately inflated judgements indicate spuriously to the student that material is understood.

When people are asked how well they will do on a test, their confidence often overshoots actual measured performance (Bandura, 1986; Fischhoff, Slovic, & Lichtenstein, 1977; Lichtenstein & Fischhoff, 1977; Metcalfe, 1986a, 1986b). Students who were about to make an error on an insight problem exhibited especially high confidence (Metcalfe, 1986b), and students who are performing poorly tend to be more overconfident than better students (Fischhoff et al., 1977). This line of research suggests two things: (1) It is important to guard against people's metacognitive feelings, since they are often misleading, and (2) the illusion of overconfidence might be offset, and performance enhanced, if students can be biased to expect that a test will be difficult, rather than easy. The risk in stressing the difficulty of learning materials in realistic educational contexts, however, is that the children may believe that the task is simply impossible, or worse, that they are not smart enough to do it, and hence give up (e.g., Bandura, 1986; Cain & Dweck, 1995; Kamins & Dweck, 1997). This risk may be especially high in the targeted minority population, given Steele's findings on the effects of stereotype threat, a predicament that can handicap members of any group about whom stereotypes exist (e.g., Steele & Aronson, 1995).

In the studies reported here we have tried to create circumstances which would allow students to make realistic judgements about their degree of learning, in several ways. One was by testing them online, as they studied. Another was by telling them at the outset, and continuing to emphasise throughout, that the task itself was extremely difficult. At the same time, we made every effort to bolster motivation, by framing the study as a game—including applause and other entertaining rewards and sounds in the programme, by calling the tutors "coaches", and telling the children that they were playing to score points, and to beat their own previous scores. Indeed, the computer pitted the child against his or her own previous performance, underlining the incremental framework advocated by Dweck (see 1990), rather than ever comparing the child's performance to that of others (which could bolster an entity framework).

THE TARGETED SCHOOL ENVIRONMENT

The experiments reported here are a follow up to those in Metcalfe (2006). Our investigations were especially targeted to a population of inner-city children in a large (1375 students) public middle school, MS 143 in New York City's South Bronx. The children in this school were at potentially high risk for school failure and a wide range of other negative behavioural and social-emotional outcomes. More than one-third of the students had literacy and academic performance scores below New York State's minimal acceptability standards. The study focused on both grades 6 and 7, the

approximate age at which highly refined study skills are becoming essential. The study programme, which will be detailed below, attempted to help the children to adopt effective, intrinsically motivated study patterns, informed by effective memory, metacognitive and planning strategies.

Our goal in conducting the experiments described below was to try to maximise the effectiveness of the teaching programme by combining the cognitive principles described above. By doing so, we hoped to demonstrate the potential such a programme has for improving learning in schools and during homework. The design does not allow us to identify the size of the role any particular factor played. We compared the computer study condition to a self study condition in which students were allowed to study as they saw fit. While we did not tell the children what to do in the self study condition, we tried to provide them with every opportunity to implement the motivational, metacognitive, and cognitive principles we implemented in the computer programme, by, for example, providing a calm quiet learning environment, materials with which to make study aids such as flashcards, and all of the words from previous sessions to encourage spaced practice.

EXPERIMENT 1

In the first experiment, we constructed a set of vocabulary items, in consultation with the children's teachers, which were deemed to be important in helping the children to understand the materials that they were studying in class, and useful in allowing them better comprehension of their science and social studies textbooks. These materials were the targets of study.

Method

Participants

The participants in this experiment were 14 children of whom eight sixth and seventh grade children completed all sessions. Six children who started the experiment did not complete it, two because they left the after-school programme and four because they were absent for too many sessions. The participants were students at a poorly performing public middle school, in New York City's South Bronx. They were treated in accordance with the guidelines of the American Psychological Association.

Design

The experiment was a 3 (study condition: computer study, self study, no study) $\times$ 2 (test condition: paper, computer) within-participant design. Study

condition was manipulated within participants, with vocabulary items assigned randomly, for each child, to one of the three conditions. Test condition was also manipulated within participants; items were tested either on paper or on the computer. This factor was included to test the possibility that there might be an advantage to computer testing in the computer study condition, and an advantage to paper testing in the self study condition. The order of conditions (paper or computer first) was a between-participants factor.

Materials and apparatus

The materials were 131 definition-word pairs; for example, Ancestor–A person from whom one is descended; an organism from which later organisms evolved. Each pair was associated with three sentences, with a blank for the target word; for example, "The mammoth is a(n) _____ of the modern elephant; Venus knew her _____(s) came from Africa and wanted to travel there to explore her roots; Bob was surprised when he found out that his _____(s) were from Norway, because this meant that he was part Norwegian."

The computer portion of the experiment was conducted using Macintosh computers. During self study, in addition to the to-be-learned materials, the children were provided with materials to make study aids, including paper, blank index cards, pens, pencils, and crayons. Self and computer study occurred in a classroom in the children's school as part of an after-school programme.

Procedure

The experiment consisted of seven sessions, which took place once a week. The first session was a pretest. Sessions 2–5 were the main learning phase of the experiment, when all of the words were introduced. Session 6 was a final review of all of the words, and Session 7 was a final test. Children participated in both the computer study and self study conditions within a given session, with the order of conditions balanced across participants.

Pretest Session 1. Participants were given a pretest to identify words they already knew. Any word answered correctly (allowing small spelling errors) was removed from the participant's pool of words, and never shown to that participant again. Spelling errors, here and in our test scoring, were assessed by a computer algorithm based on the amount of overlap between the set of letters used in the subject's answer and the set of letters in the correct answer, and also takes letter order into account. Using this algorithm, the program produced a degree of correctness score from 0 to 100. Small errors were those that had a numerical score between 85 and 99.

During the pretest, the 131 definitions were presented one at a time, in random order. The child's task was to type in the corresponding word and press return. (Although there was some variation in speed, none of the children had problems typing.) Questions remained on the screen for a maximum of 30 s, after which, if the child had not made a response, the question disappeared, and the response was scored as being incorrect. Once a question had either been answered or had timed out, a "Next" button was shown, and when it was pressed, a new question appeared. The entire pretest was conducted on the computer. The pretest continued until the participant got 120 items wrong (which all participants did).

Training (Sessions 2–5).   Sessions lasted for 35 min of self study and 35 min of computer study. During each session, half of the students did computer study and half did self study for 35 min, and then they switched places. Twenty new words were presented during each session, 10 for self study and 10 for computer study.

During self study, the students were given sheets of paper containing each of the 10 new word/definition pairs and their accompanying sentences. The word/definition pair was on one side of the paper, while the three contextual sentences (which included a blank for the target word in question, accompanied by each target word) were on the other side. Students, with their study aids in hand, were allowed to study however they saw fit. At the end of each session all of the study materials the participant had been given or created were put into a folder. At the start of the next session they were given their folder, so that if they chose to, they could study words from all of the preceding sessions, using the materials we had provided and the study aids they had created. Although the children in this condition had the opportunity to space their practice, they were not obligated to do so.

During self study, the experimenters made every effort to ensure that the students paid attention: The classroom environment was quiet and calm, the desks were separated so that the children could not distract each other, and when it seemed to be necessary, the tutors approached the children one-on-one and encouraged them to stay on task. We did this in an effort to equate (or at least increase the equivalence of) motivation between self and computer study, based on the expectation that the computer condition would naturally be more engaging than the self study condition, in part because of the cognitive principles we implemented (e.g., testing), and in part because of the simple fact that it was interactive. It seemed clear that the children behaved during self study in the same way they behaved during class. Because of the small group of students and the presence of multiple tutors, the quality of the self study conditions we provided, when compared to the situations the children normally studied in at home and in school, bordered on being unrealistic.

Computer study during Sessions 2–5 consisted of three main phases. The first phase was a test. During the test, all of the words that had been presented in previous sessions were tested. There was no test in Session 2, because it was the first real session; in Session 3, the 10 words from Session 2 were tested; in Session 4, the 20 words from Sessions 2 and 3 were tested; in Session 5, all 30 words from the previous sessions were tested. During the test at the beginning of each session, the definition and one of the three accompanying sentences were presented, and participants were asked to type in the target word. Each response had to be at least three letters long. Correct responses that were spelled perfectly were followed by a recording of applause played by the computer. If the participant did not respond after 30 seconds the response was automatically coded as incorrect and the computer gave a 'bad' – but slightly funny – beep, followed by the correct word. Responses that were close, but not spelled correctly, were followed by a recording saying, "Close, it's" followed by a recording of the word. The word was then shown on screen. If the response was incorrect a bad beep was played, followed by a recording of the word, and the word was shown on the screen. After the feedback, a "Next" button appeared.

The second phase of computer study was new item presentation. Each of the day's new words was shown individually, along with its definition and an accompanying sentence. Each target/definition/sentence compound was presented for 10 s.

The third phase consumed the majority of the session. In this phase, participants studied the current session's 10 new words. The definition and one of the three accompanying sentences was shown, and the participant was asked to type in the target. The programme cycled through the sentences. If the response was correct, a recording of the word was played, the word was shown on the screen, and the "Next" button appeared. When the "Next" button was pressed, the next definition and accompanying sentence were presented. If the response was incorrect, a bad beep was played followed by a recording of the word, and then the "Next" button appeared. When the "Next" button was pressed, the programme presented the same word again, and continued to do so, using the same procedure, until a correct response was given. From the fourth incorrect guess onwards, the correct answer was shown onscreen at the same time as it was spoken. For close responses that were not spelled correctly, instead of the bad beep, a recording saying, "Close, it's" was played followed by the word, and the word was shown on the screen (regardless of which trial it was).

The definitions were presented repeatedly, in cycles, such that all of the items were first presented in random order, and then, once that cycle was complete, they were all presented again in a rerandomised order. For example, if there had been only four items instead of ten, the order of the first three cycles might have been 2,1,3,4 . . . 4,2,3,1 . . . 2,1,3,4.

As study progressed, words that were considered temporarily learned were removed from the pool of cycling words to be presented. To be considered temporarily learned a word had to be answered correctly, on the first try, on two separate occasions. In this way, the focus was put on the words that the participant had not learned, and time and effort were not wasted on words they already knew (see Kornell & Metcalfe, 2006; Metcalfe & Kornell, 2005, for more details concerning the efficacy of this strategy).

Study continued until all of the words had been answered correctly twice. At that point, the 10 words that the participant had been studying were added back into the study cycle, along with all of the words that had been answered incorrectly on the pretest, in random order. From this point on, no words were removed from the cycle, and study continued until the 35-min session ended.

At the end of the session, all of the words that the participant had answered correctly on the first try twice, or had answered correctly on the pretest at the start of the session, were displayed on the screen as positive reinforcement, and the participants were told that these were the words they had learned today.

Before beginning the first training session (Session 2), the children were given verbal instructions describing the experiment. They were told that over the course of the next few sessions they would be learning the words that they had seen during the pretest. It was explained that each day they would be given the chance to study words from the previous sessions, as well as new words. They were told that they would have a folder for self study, which would contain all of the materials they had worked on, so that when they started a new session they would be able to use whatever materials they had created in previous sessions. They were also told that the materials they would be learning would be words they would need to know for their classes in school.

Final review (Session 6).   During the final review no new words were presented. For self study, the participants were simply given the folder with the four sheets of words, plus any study aids they had made in the previous weeks. For computer study, the session began with a test on all of the words. Then study commenced on all of the words that were not answered correctly on the test, in the same way as in previous weeks. If and when all of the words had been taken out of the study cycle, all 40 words were added back into the cycle, and study continued with no words being removed, until the 35-min session was complete.

Final test (Session 7).   During the final test all 120 words were tested. Sixty were tested on the computer and sixty were tested on paper, with half of the items from each condition (computer study, self study, no study) being

included on each test. On the paper test, the 60 definitions were shown with blank spaces where the words were to be filled in. On the computer test, the definitions were shown one by one, and the participant was asked to type in the response and press return. No feedback was given on either test.

Results and discussion

The data were analysed using an ANOVA with two factors: type of study (computer study, self study, no study) and type of test (computer, paper). The dependent variable was performance on the final test. Performance was scored in two ways, leniently and strictly. With strict scoring, only answers spelled correctly were considered correct, whereas with lenient scoring close answers that were spelled incorrectly were considered correct. The two methods of scoring resulted in the same patterns of significance across all three experiments, so we present only the lenient data. To score the children's responses on the paper test, after the data were collected, the experimenters entered the children's responses on the paper test into a computer program, which scored them, so that the same scoring algorithm was used for both types of test.

The results are shown in Figure 1. There was a significant effect of type of study, $F(2, 14) = 42.19$, $p < .0001$, $\eta_p^2 = .86$ (effect size was computed using
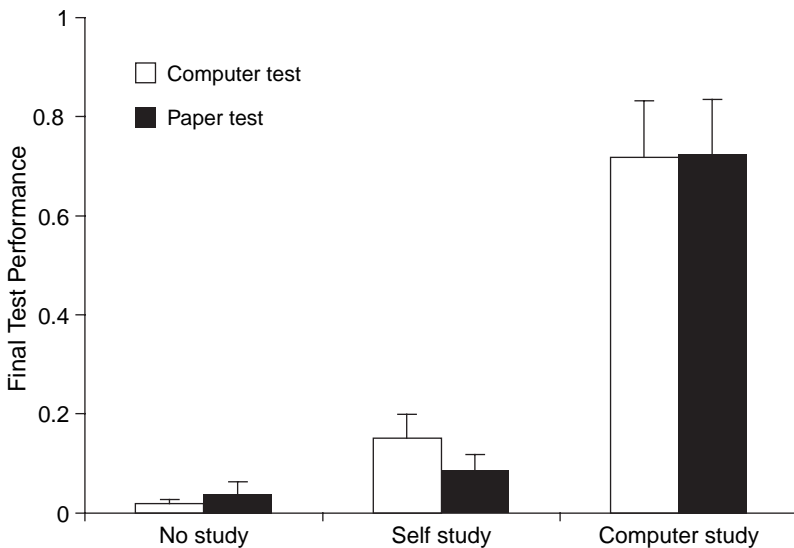

Figure 1.   Experiment 1, children's final test performance in the computer study, self study, and no study conditions. Scores are presented for both the computer test and paper test. Error bars represent standard errors.

partial eta squared). The main effect of type of test was not significant, nor was the interaction. Post hoc Tukey tests showed that performance in the computer study was significantly better than self and no study conditions. The self study condition was not significantly better than the no study condition.

The day on which a word was first studied had a significant effect on final test performance. An analysis of words studied in the computer and self conditions (the no study condition was excluded because the words were not studied on any day) showed that the earlier an item was introduced, the greater its chance of being recalled on the final test, $F(3, 21) = 3.16$, $p < .05$, $\eta_p^2 = .31$. This effect can be explained by the fact that words presented in earlier sessions were restudied in later sessions; as a result, they were presented more times and with greater spacing than words introduced in later sessions.

The results were overwhelmingly favourable for the cognitive-science based programme. Encouraged, we designed Experiment 2 to investigate the efficacy of a similar programme, intended to teach English vocabulary to students who were native Spanish speakers learning English as a second language, in the same at-risk school.

EXPERIMENT 2

Besides the fact that the materials were Spanish–English translations, the main difference between this experiment and Experiment 1 was that the children were actively trying to learn the targeted words in their classrooms. The materials were relatively easy, and were, for the most part, high frequency words referring to common everyday objects.

Method

Participants

The participants were 25 sixth and seventh grade children from the same school as Experiment 1. Seven children who started the experiment but did not finish, four because they left the after-school programme and three because they were absent for too many sessions, leaving eighteen children who completed all sessions. These children were native Spanish speakers, with varying levels of English competence, who were in their school's English as a Second Language (ESL) programme.

Design

The design was a within-participants design with two conditions: computer study and self study.

Materials and apparatus

The materials were 242 Spanish–English translations, e.g., flecha–arrow. In cases in which multiple Spanish synonyms seemed appropriate for an English word, more than one Spanish word was shown, e.g., escritorio, bufete–desk. Pictures of the items were taken from the Snodgrass picture set (Snodgrass & Vanderwart, 1980).

The computer portion of the experiment was conducted using Macintosh computers. During self study, paper, blank index cards, pens, pencils, and crayons were available for the children to make study aids. Self and computer study occurred in a classroom in the children's school as part of an after-school programme.

Procedure

The experiment consisted of six sessions, which took place once a week. The first session was a pretest. Sessions 2–5 were the main learning phase of the experiment, when all of the words were introduced. Session 6 was a final test. The instructions and conditions of self study were similar to Experiment 1.

Pretest Session 1.   During the pretest, each of the 242 Spanish words was presented one at a time on the computer. The child's task was to type in the English equivalent and press return. No feedback was given about the accuracy of the responses. Questions could be on the screen for a maximum of 30 s, after which the question disappeared, the response was scored as incorrect, and a button labelled "Listo/a" ("Next") was shown. The reason for the pretest was to identify words the participants already knew. Any word answered correctly (with correct spelling) was taken out of the participant's pool of words, and never shown to that participant again.

Training (Sessions 2–5).   Training sessions were divided into two parts, self study and computer study, each of which lasted 30 min. During each session, half of the students engaged in computer study and half in self study, and then they switched places. Forty new words were presented during each session, twenty for self study and twenty for computer study.

During self study, the students were given sheets of paper with each word in both Spanish and English, as well as a picture of the object. They were also given access to blank paper, index cards, pencils, crayons, and pens, so that they could make flashcards or other study aids. As in Experiment 1, participants were allowed to study however they saw fit. At the end of each session all of the study materials the participant had been given or created were put into a folder. At the start of the next session they were given their

folder, so that if they chose to, they could study words from all of the preceding sessions, using the materials we had provided and the study aids they had created.

Computer study during Sessions 2–5 consisted of three main phases. The first phase was a pretest (distinct from the session one pretest). During the pretest the "unlearned" translations were presented. The "unlearned" translations included all items that had been presented in previous sessions but had not been answered correctly on subsequent pretests. Being answered correctly on a pretest was the only way for a word to be considered "learned". The pretest did not include the current session's new items; thus, there was no pretest during Session 2, because there were no words that had already been presented. In Session 3, all 20 words from Session 2 were included on the pretest. For example, if the participant answered 7 items correctly on the pretest in Session 3, the remaining 13 words would be pretested again in Session 4 (along with the 20 new words from Session 3).

During the pretest, each Spanish word was presented individually, and participants were asked to type in the English translation. Correct responses that were spelled perfectly were followed by a rewarding "ding" and then a recording of the word being spoken. If the response was close, but not spelled correctly, a recording saying, "Close, it's" was played, followed by a recording of the word, and the word was shown on screen (all of the instructions were in Spanish except for "Close, it's"). If the response was incorrect the recording of the word was played, and the word was shown on the screen. After this feedback, the "Next" button appeared.

The second phase of computer study was initial presentation. Each pair of words was presented once, one pair at a time. First the Spanish word was shown. After 1 s the English word was shown, and then after another 1 s a picture of the item was shown and the word was spoken aloud simultaneously. One second later, the screen cleared and the next word appeared.

The third phase of computer study, during which the session's 20 new words were presented and studied, consumed the majority of the session. The procedure for each presentation during study was as follows: Either the Spanish word or a picture of the item was presented, and the participant was asked to enter the English response. If the response was correct, a beep was played followed by a recording of the word, and the "Next" button appeared. When the "Next" button was pressed, a new word was presented. If the response was incorrect, a bad beep was played followed by a recording of the word, and then the next button appeared. When the "Next" button was pressed, the programme presented the same word again, and continued to do so, using the same procedure, until a correct response was given. From the fourth incorrect guess onwards, the correct answer was shown onscreen at the same time as it was spoken. For close responses that were not spelled

correctly, instead of the bad beep, a recording saying, "Close, it's" was played followed by the word, and the word was shown on the screen.

The words were presented repeatedly, in cycles, such that all of the words were presented in random order, and then they were all presented again in a rerandomised order, and so on. For example, if there had been only 4 new items instead of 20, the order of the first three cycles might have been 2,1,3,4 ... 4,2,3,1 ... 2,1,3,4. As study progressed, words that were considered temporarily learned were removed from the cycle of words to be presented. In order to be considered temporarily learned a word had to be answered correctly, on the first try, on two separate occasions. In this way, the focus was put on the words that the participant had not learned. To insure that the participant could answer based on both verbal and pictorial cues, once a correct answer had been given to a picture, the first presentation was always a word from then on, and vice versa.

Study continued until all of the words had been answered correctly twice (or the session ended). At that point, the 20 new words that the participant had just been studying, along with all of the words that had been answered incorrectly on the pretest, were added to the list in random order and study continued. Again, if a word was answered correctly on the first try twice it was removed from the cycle. When all of the words had been removed again, this larger set of words was once again added back into the list, and study continued with none of the words being removed.

At the end of each computer session, the participant was shown two animated cars racing across a track. One represented their memory performance this session, and one their performance in the previous session. This was used as a way to encourage participants to try to do better each session than they had done previously, promoting learning and effort instead of absolute performance.

Final test (Session 6).   During the final test 80 self and 80 computer words were presented in random order. Since there was no difference between testing on paper and on the computer in Experiment 1, the entire test was conducted on the computer. The Spanish word was shown, and participants were asked to type in the English word and hit return. Correct answers were followed by a "ding" and then the sound of the word being spoken. Incorrectly spelled answers that were close to correct were followed by a recording that said, "Close, it's" and then a recording of the word. Incorrect answers were followed by a bad beep and then a recording of the word. During the final test, the word was not shown on the screen.

After the 160 words that had been studied were tested, a set of words that had not been studied was tested. The number of unstudied words depended

on the number of words the participant had answered correctly on the pretest in Session 1. Because some children had very few (or in one case, zero) unstudied words, we decided to test them at the end of the experiment and did not include them in the data analysis. Test performance for the words that were not studied was lower than for the two other conditions, as expected. The mean accuracy was only 0.14. A score could not be computed for one child who had no words left to be assigned to the no study condition at the end of the pretest.

Results and discussion

As in Experiment 1, there were no differences between strict and lenient scores so only the leniently scored data are presented here. Computer study led to improved performance on the final test, as Figure 2 shows. Performance on the final test was significantly better in the computer study condition ($M = 0.41$) than in the self study condition ($M = 0.29$), $F(1, 17) = 13.61$, $p < .01$, $\eta_p^2 = .44$. Final test performance was also significantly better for items that were first presented in the earlier sessions than items introduced later, $F(3, 51) = 5.08$, $p < .01$, $\eta_p^2 = .23$. As with the first experiment, the results were highly favourable for the cognitively enhanced computer study.
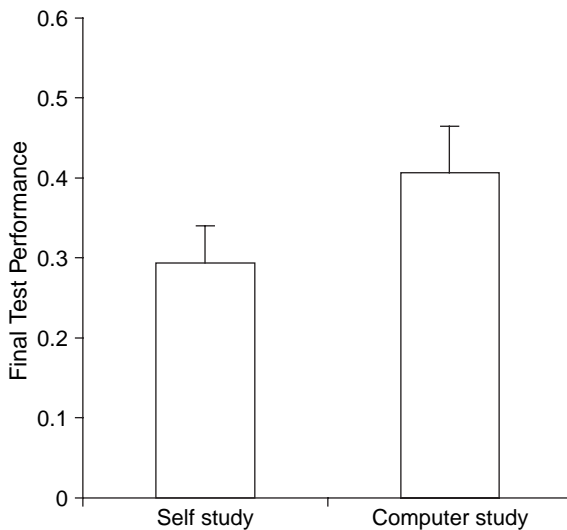

Figure 2. Experiment 2, children's final test performance for the self study and computer study conditions. Error bars represent standard errors.

EXPERIMENT 3

In Experiment 3 we used Columbia University undergraduates as participants. Unlike the young, at-risk children, we expected the undergraduates to have well-developed study skills and to be highly motivated. We expected, though, that even with the highly motivated and skilled Columbia University students, there would still be an advantage for the computer-based study programme, because it circumvented metacognitive illusions, required self generation, and implemented spacing, factors that are difficult to implement without the assistance either of a computer or another person versed in such techniques.

Method

Participants

The participants were 14 Columbia University students who did not speak Spanish. One participant began the experiment but did not complete all of the sessions, leaving 13 complete participants.

Materials and apparatus

The materials were the same 242 Spanish–English translations used in Experiment 2. In Experiment 2, in some cases, multiple Spanish synonyms were used as a single cue. In this experiment, however, the task was to type in the Spanish translation of an English cue. To avoid requiring participants to type in multiple synonyms, we removed all but one of the Spanish synonyms in such cases, for example, changing the pair escritorio, bufete–desk to escritorio–desk. The pictures of the items, which were the same as in Experiment 2, were taken from the Snodgrass pictures (Snodgrass & Vanderwart, 1980).

The computer portion of the experiment was conducted using Macintosh computers. During self study, paper, blank index cards, pens, pencils, and crayons were available for the participants to make study aids. Self and computer study occurred in a laboratory at Columbia University. Like the children in Experiment 2, the students in this experiment were run as groups in a room together.

Design and procedure

The design and procedure of Experiment 3 were the same as Experiment 2, with the following exceptions: In Experiment 3, participants were asked to type in Spanish translations of English words, whereas in Experiment 2 they were asked to type in the English translations of Spanish words. In both experiments, however, the participant's own language was used as the cue

and they were asked to type in the language they were learning. Also, the instructions in Experiment 3 were in English instead of Spanish.

Results

The results of Experiment 3, shown in Figure 3, mirrored those of Experiment 2. Performance on the final test was significantly better in the computer study condition ($M = 0.76$) than in the self study condition ($M = 0.67$), $F(1, 12) = 4.96$, $p < .05$, $\eta_p^2 = .29$. Final test performance was also better for words first presented in earlier sessions than words introduced later, $F(3, 36) = 52.52$, $p < .0001$, $\eta_p^2 = .81$. Test performance for words that were not studied was 0.10.

CONCLUSION

In three experiments, we tested the effectiveness of a computer-based learning programme that used principles of cognitive science, including elaborative processing, generation, transfer-appropriate processing, spacing, and metacognitive and motivational techniques. All three experiments showed that the computer-based programme led to significantly more learning than self study. The boost in performance was present for at-risk
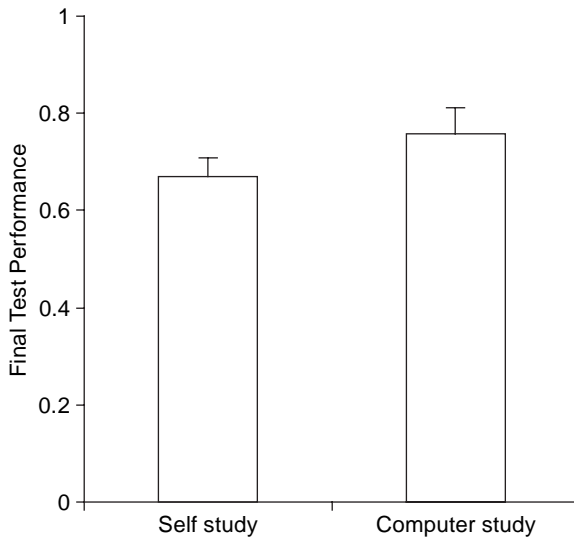

Figure 3. Experiment 3, college students' final test performance for the self study and computer study conditions. Error bars represent standard errors.

middle school students (Experiments 1 and 2) and for the more experienced college students (Experiment 3).

Our research here was based on the work of previous researchers who have discovered a number of strategies (e.g. generation, testing, and spacing) that have benefited learning in the laboratory. The application of these strategies in the classroom has lagged behind the laboratory research. However, there has been a recent surge in interest in both isolating principles of learning that are likely to have beneficial effects in classroom settings, and in actually implementing such principles to improve children's learning. The efficacy of the translation of some of these components is beginning to emerge. For example, in a series of recent papers, (Chan, McDermott, & Roediger, in press; Kuo & Hirshman, 1996; McDaniel, Anderson, Derbish, & Morrisette, 2007 this issue; Roediger & Karpicke, 2006; Roediger, McDaniel, & McDermott, 2006) it has been shown that not only does testing improve memory more than does additional studying, but that this "testing effect" occurs in the classroom as well as in the lab. We used multiple tests in our programmes, and so it is reassuring to see that this was well justified: When the "testing effect" that we made use of in our composite study has been isolated in classroom situations, it does, indeed, improve learning.

The generation effect, long established in the laboratory, and aggressively used in the computer-based study programmes that we employed here, has a more chequered recent history in classroom situations. Metcalfe and Kornell (in press), and deWinstanley's group (deWinstanley & Bjork, 2004; deWinstanley, Bjork, & Bjork, 1996) have found that the generation effect is not always in evidence when performance on the generate items is compared to a control condition in which items were presented to the students. The reason for this apparent failure of a well-replicated laboratory finding may not be that the generation effect itself is not robust, however, but rather that it may be too robust. It may be very easy to get students, in a classroom or study situation, to mentally generate, even when they are nominally in the control condition. They may simply generate spontaneously or they may generate because they have learned that it is advantageous to their performance. When they do not generate, however, as Metcalfe and Kornell's (in press) follow-up experiments show, performance is impaired both in the lab and in the classroom setting. Thus, evidence is accumulating that the strategy that we used, of forcing the children to generate, was, indeed a good one. One might wonder whether and to what extent the effects of generating are mitigated by the fact that the children sometimes generate errors. In the programmes reported here, many errors were, indeed, generated, though corrective feedback was given.

Finally, while the children were free to space their practice in the self study condition, they were forced to do so in the computer controlled programme, and presumably spacing was used more rigorously in the latter than the former

condition in the studies presented here. While the laboratory evidence is overwhelming that spacing is beneficial in this kind of learning, the data from the classroom for the advantages of spacing are only now beginning to emerge. Interestingly, while the memory effects for factual material are holding up, Pashler, Rohrer, and Cepeda (2006) have recently reported that in certain practical domains—such as in the acquisition of visual-spatial skills, or in certain categorisation tasks—spacing is unimportant (but see Kornell & Bjork, in press, for evidence that spacing is effective in a categorisation task). Future research, directed at improving teaching in particular domains and for particular targeted skills and knowledge bases, may reveal important qualifications concerning the generality of spacing effects. The tasks that we explored in our study were classic verbal acquisitions tasks, and as such we expected spacing to be important. However, by taking aim at the materials and concepts that children need to learn in school situations, we may reveal important limitations to what have, almost universally, been accepted as fundamental principals of learning such as the advantage of spaced practice.

The primary goal of this computer-based study was to take effective cognitive and metacognitive strategies out of the laboratory, and implement them in a real learning situation, and to compare the outcome with students' natural learning strategies. The findings show that young learners are often prone to spontaneously using *ineffective* study strategies, and that even accomplished college students, who have a great deal of practice studying effectively, would benefit from direct cognitive and metacognitive learning programmes. The current programme is only a small step towards bridging the gap between laboratory testing and individual learning, but these data show that laboratory research on learning principles can be valuable in enhancing learning in classroom situations. We conclude that learning programmes that employ motivational, cognitive, and metacognitive techniques together are effective learning tools. In the future, it would be of interest to further investigate the magnitude of the role each strategy played in improving learning. Such a research strategy may serve not only to consolidate our confidence about the procedures that we have used, when those procedures survive translation into a classroom setting and are appropriate, but it may also, serendipitously, reveal that some of those supposedly well-established principles are not what they seem, despite a great deal of laboratory research.

REFERENCES

Baddeley, A. D. (1978). The trouble with levels: A re-examination of Craik and Lockhart's framework for memory research. *Psychological Review, 85*, 139–152.

Baddeley, A. D., & Longman, D. J. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics, 21*, 627–635.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*(4), 566–577.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

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

Bjork, R. A. (1970). Repetition and rehearsal mechanisms in models of short-term memory. In D. A. Norman (Ed.), *Models of memory* (pp. 307–330). New York: Academic Press.

Bjork, R. A. (1975). Short-term storage: The ordered output of a central processor. In F. Restle, R. M. Shiffrin, N. J. Castellan, H. R. Lindman, & D. B. Pisoni (Eds.), *Cognitive theory* (pp. 151–171). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bjork, R. A. (1988). Retrieval practice. In M. Gruneberg & P. E. Morris (Eds.), *Practical aspects of memory: Current research and issues, Vol. 1: Memory in everyday life* (pp. 396–401). New York: Wiley.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. J. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bransford, J. D., & Johnson, M. K. (1973). Considerations of some problems of comprehension. In W. Chase (Ed.), *Visual information processing* (pp. 383–438). New York: Academic Press.

Cain, K. M., & Dweck, C. S. (1995). The relation between motivational patterns and achievement cognitions through the elementary school years. *Merrill-Palmer Quarterly*, *41*, 25–52.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*, 633–642.

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

Cermak, L. S., & Craik, F. I. M. (1979). *Levels of processing in human memory*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cermak, L. S., Naus, M. J., & Reale, L. (1976). Rehearsal strategies of alcoholic Korsakoff patients. *Brain and Language*, *3*, 375–385.

Cermak, L. S., & Reale, L. (1978). Depth of processing and retention of words by alcoholic Korsakoff patients. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 165–174.

Chan, J. C. K., McDermott, K. B., & Roediger, H. (in press). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235.

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*, 945–955.

DeWinstanley, P. A., Bjork, E. L., & Bjork, R. A. (1996). Generation effects and the lack thereof: The role of transfer-appropriate processing. *Memory*, *4*, 31–48.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20,* 374–380.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545–565.

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

Dweck, C. S. (1990). Self-theories and goals: Their role in motivation, personality, and development. In R. A. Dienstbier (Ed.), *Nebraska symposium on motivation* (Vol. 38, pp. 199–235). Lincoln, NE: University of Nebraska Press.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564.

Fisher, R. P., & Craik, F. I. M. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 701–711.

Gardiner, J. M., & Hampton, J. A. (1985). Semantic memory and the generation effect: Some tests of the lexical activation hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 732–741.

Gardiner, J. M., & Rowley, J. M. (1984). A generation effect with numbers rather than words. *Memory & Cognition, 12,* 443–445.

Geiselman, R. E., & Bjork, R. A. (1980). Primary and secondary rehearsal in imagined voices: Differential effects on recognition. *Cognitive Psychology, 12,* 188–205.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15,* 1–16.

Glenberg, A. M. (1979). Component levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7,* 95–112.

Graf, P. (1980). Two consequences of generating: Increased inter- and intraword organization of sentences. *Journal of Verbal Learning and Verbal Behavior, 19,* 316–327.

Hannon, B., & Craik, F. I. M. (2001). Encoding specificity revisited: The role of semantics. *Canadian Journal of Experimental Psychology, 55,* 231–243.

Hintzman, D. L. (1987). Recognition and recall in MINERVA2: Analysis of the "recognition-failure" paradigm. In P. Morris (Ed.), *Modeling cognition* (pp. 215–229). New York: Wiley.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649–667.

Jacoby, L. L., & Kelley, C. M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin, 13,* 314–336.

Johns, E. E., & Swanson, L. G. (1988). The generation effect with nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 180–190.

Kamins, M. L., & Dweck, C. S. (1997, May). *The effect of praise on children's coping with setbacks.* Poster presented at the ninth annual meeting of the American Psychological Society, Washington, DC.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32,* 1–24.

Koriat, A. (1994). Memory's knowledge of its own knowledge: The accessibility account of the feeling of knowing. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 115–135). Cambridge, MA: MIT Press.

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

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

Kornell, N., & Bjork, R. A. (in press). The promise and perils of self-regulated learning. *Psychonomic Bulletin and Review.*

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 609–622.

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109*, 451–464.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159–183.

Lockhart, R. S., Craik, F. I. M., & Jacoby, L. L. (1976). Depth of processing, recognition, and recall: Some aspects of a general memory system. In J. Brown (Ed.), *Recall and recognition* (pp. 75–102). London: Wiley.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, M. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 494–513.

Metcalfe, J. (1982). A composite holographic associative recall model. *Psychological Review, 89*, 627–661.

Metcalfe, J. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review, 92*, 1–38.

Metcalfe, J. (1986a). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 288–294.

Metcalfe, J. (1986b). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 623–634.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review, 100*, 3–22.

Metcalfe, J. (1995). Distortions in human memory. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 321–322). Cambridge, MA: MIT Press.

Metcalfe, J. (1997). Predicting syndromes of amnesia from a Composite Holographic Associative Recall/Recognition Model (CHARM). *Memory, 5*, 233–253.

Metcalfe, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics? *Personality and Social Psychology Review, 2*, 100–110.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*, 349–363.

Metcalfe, J. (2006). Principles of cognitive science in education. *APS Observer, 19*, 27–38.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463–477.

Metcalfe, J., & Kornell, N. (in press). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin and Review.*

Metcalfe, J., & Murdock, B. B., Jr. (1981). An encoding and retrieval model of single-trial free recall. *Journal of Verbal Learning and Verbal Behavior, 20*, 161–189.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 851–861.

Murdock, B. B., Jr., & Metcalfe, J. (1978). Controlled rehearsal in single-trial free recall. *Journal of Verbal Learning and Verbal Behavior, 17*, 309–327.

Nairne, J. S., & Widner, R. L. (1987). Generation effects with nonwords: The role of test appropriateness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 164–171.

Nelson, T. O. (1977). Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 151–172.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs, 76*, 28.

Pashler, H., Rohrer, D., & Cepeda, N. J. (2006). Temporal spacing and learning. *Observer, 19*, 19.

Peynircioglu, Z. F. (1989). The generation effect with pictures and nonsense figures. *Acta Psychologica, 70*, 153–160.

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.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology, 19*, 90–138.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 435–451.

Roediger, H. L., Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels of processing framework. *Memory, 10*, 319–322.

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

Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (2006). Test-enhanced learning. *APS Observer, 19*, 28–38.

Rohrer, D., Taylor, K., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology, 19*, 361–374.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89*, 63–77.

Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1074–1083.

Shaughnessy, J. J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Verbal Learning and Verbal Behavior, 20*, 216–230.

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 907–912.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 174–215.

Son, L. K. (2005). Metacognitive control: Children's short-term versus long-term study strategies. *Journal of General Psychology, 132*, 347–363.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage model. *Memory and Cognition, 33*, 1116–1129.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811.

Tulving, E. (1983). *Elements of episodic memory.* Oxford, UK: Oxford University Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352–373.

Weaver, C. A., III, & Kelemen, W. L. (1997). Judgments of learning at delays: Shifts in response patterns or increased metamemory accuracy? *Psychological Science, 8*, 318–321.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15*, 41–44.