Psychonomic Bulletin & Review
2007, 14 (2), 225-229

Principles of cognitive science in education: The effects of generation, errors, and feedback

Janet Metcalfe
Columbia University, New York, New York

and

Nate Kornell
University of California, Los Angeles, California

Principles of cognitive science hold the promise of helping children to study more effectively, yet they do not always make successful transitions from the laboratory to applied settings and have rarely been tested in such settings. For example, self-generation of answers to questions should help children to remember. But what if children cannot generate anything? And what if they make an error? Do these deviations from the laboratory norm of perfect generation hurt, and, if so, do they hurt enough that one should, in practice, spurn generation? Can feedback compensate, or are errors catastrophic? The studies reviewed here address three interlocking questions in an effort to better implement a computer-based study program to help children learn: (1) Does generation help? (2) Do errors hurt if they are corrected? And (3) what is the effect of feedback? The answers to these questions are: Yes, generation helps; no, surprisingly, errors that are corrected do not hurt; and, finally, feedback is beneficial in verbal learning. These answers may help put cognitive scientists in a better position to put their well-established principles in the service of children's learning.

In the 40 years since the cognitive revolution took hold in American psychology, many advances have been made in our understanding of how people think and remember, as well as in what people know about what they know—that is, their metacognition. These advances have come about largely through a plethora of exciting laboratory experiments. As a result, there is a variety of memory phenomena that, in the laboratory, are both highly replicable and well understood. Deep encoding, spaced practice, test-enhanced learning, the generation effect, the encoding specificity principle, the effects of difficult retrieval, the beneficial effects of spacing, judgments of learning, and over- and underconfidence biases are all buzzwords within cognitive psychology that have implications for enhancing learning in a classroom situation. Our increasing understanding of people's metacognitive biases and distortions may be particularly important. But applying these principles in a real-world setting frequently gives rise to unexpected issues that may modulate their effectiveness.

One of the most clearly applicable sets of findings are the biases and illusions in people's metacognition. People frequently believe they have learned something when in fact they have not (Bjork, 1999; Metcalfe, 1998). Given such metacognitive failures, learners may be in a poor position to remedy their own faulty study habits. Furthermore, there are many real-world situations in which the task itself destroys our metacognition. For example, when people study for a foreign language vocabulary test, the fact that the textbook presents the to-be-remembered word side by side with its translation would seem to be innocuous. But it produces exactly the condition, shown by Kelley and Jacoby (1996), that evokes overconfidence. Because people's choice of whether to study or not is directly related to whether they think they know or not (Metcalfe & Finn, in press), they will decline study because they think, wrongly, that they already know.

One way to circumvent metacognitive illusions is to encourage people to test themselves, which has benefits both because it allows for more accurate metacognition (Dunlosky, Hertzog, Kennedy, & Thiede, 2005) and because the effects of the test are themselves beneficial (Glover, 1989; Hogan & Kintsch, 1971; Roediger & Karpicke, 2006), perhaps because the person generates the answers actively. But without the help of a teacher, a tutor, a parent, or a computer, self-testing can be logistically problematic, if the learner thinks of doing it at all. Educators and psychologists need to devise ways to induce people to do this rather than allow them to automatically fall into a dysfunctional, illusory metacognitive state. On a related note, generating the answer rather than simply reading it or having it presented gives rise to a well-documented beneficial memory effect (Hirshman & Bjork, 1988; Slamecka & Graf, 1978). But in the course of book learning and listening to lectures, people are

J. Metcalfe, jm348@columbia.edu

 Copyright 2007 Psychonomic Society, Inc.

often put into passive reading or presentation situations. Furthermore, when people are allowed to decide when to stop studying, their memory performance can be worse than when the experimenter controls their timing (Kornell & Bjork, 2007). People do not realize when extra study time will help (Koriat, 1997). They may not spontaneously space their learning (see Benjamin & Bird, 2006), although there is abundant evidence that they should do so (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006).

For these reasons, computer-based study programs, which eliminate some or all of these problems and biases, show great promise in helping students to learn. But the question of how best to structure a computer-based study program remains. To begin investigating this issue, we devised a program to assist vocabulary learning based on principles of cognitive science. This project will be described below, before we elaborate on some of the problems and issues that we encountered in the effort.

Background Research: The Bronx Project

We began our work, in collaboration with teachers and principals, in an at-risk inner-city public middle school in New York City's South Bronx. The Grade 6 children who were our volunteers, who had very low literacy and academic performance scores, were at potentially high risk for school failure and a wide range of other disadvantageous behavioral and social-emotional outcomes. Together with their teachers, we constructed a list of vocabulary words that the children needed to know to be able to read their textbooks and to understand the materials on evaluative tests. We developed a computer-assisted study program that we hoped would allow metacognitive illusions to be overcome and implement many of the principles that we and other researchers had studied in the lab: multimodal presentation in case there were reading difficulties, spaced practice, repeated quizzing, and study time allocation based on ongoing evaluations of the children's performance. Rather than being allowed to simply read passively or, indeed, ignore the to-be-learned material, the children had to generate the answers. The program provided applause when they were right and corrective feedback when they were wrong. Repeated study and testing was conducted over a course of 7–8 weeks.

Each day, half of the children started on a randomly selected subset of the to-be-learned vocabulary items in a *self-study* condition, in which they had all of the standard study aids—flashcards, colored pens, papers, and so on—that they would ideally have in a quiet study room. The other half of the children started on the computer-based program. At the end of 35 min of studying, the children switched. Thus, each child served as his or her own control. At the end of the 7th week, all of the studied vocabulary, plus a subset that had not been studied at all, was tested.

The results were highly favorable for the cognitive-science-based study program. In the first study we conducted (see Metcalfe, 2006), performance in the computer condition was more than 400% better than that in the self-study condition (which was very low), and in a second, more refined, 7-week study (Metcalfe, Kornell, & Son, in press) it was more than 600% better than in the self-study condition. These results were encouraging to us, and we replicated them twice more—first with children from the same community learning English as a second language, and then with Columbia University students (Metcalfe et al., in press). There was a significant benefit for the computer-based study program with all of these groups.

Questions That Arose From the Bronx Project

Although we were encouraged, we had also faced a number of conundrums in developing the programs. We had had to make many decisions without sufficient knowledge. The results showed that the combination of interventions was effective but did not allow us to identify exactly what was responsible or whether some procedures we had included might actually be harming learning.

For example, influenced by the many studies that had shown such effects, we believed that self-generation of the answers was crucial. Therefore, we had implemented a self-generation procedure even though we also knew that the children would inevitably make mistakes. We did not know how detrimental the mistakes would be. Perhaps the learners would remember the mistakes, and the misinformation from them would become so embedded that the use of a generation procedure would not be warranted. Perhaps we should have had the children generate the answers only when they were very sure and likely to be correct, rather than all of the time. In that case, we might have been able to reap the benefits of the generation procedure without suffering the (supposed) impairments due to the production of mistakes.

To offset the anticipated problem that the children might learn the errors, we had given feedback and never let an error stand uncorrected. This seemed reasonable, but we did not know what the effects of the feedback would be. Feedback has been addressed in the domain of educational psychology (for reviews, see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Butler & Winne, 1995), but in the cognitive psychology literature we found a shortage of articles (but see Pashler, Cepeda, Wixted, & Rohrer, 2005, for new research). Despite our intuition that feedback would help, there were data showing that too much feedback impairs rather than helps motor skill learning (Bjork, 1999; Schmidt & Bjork, 1992). Were we far enough removed from this kind of learning that feedback would be beneficial?

To address these uncertainties, we conducted several laboratory-based experiments, based on a single design, on the effects of free versus forced generation, errors, and feedback (see Metcalfe & Kornell, 2007). It was our hope that by answering these questions, we would be able to revise the original study program and do an even better job of enhancing children's learning.

EXPERIMENT

Method

Design. Using the Bronx Project, described above, as the model for the present experiment, we designed a 3 × 2 factorial experi-

ment (Metcalfe & Kornell, 2007). The first factor was generation (free generation [i.e., generate only if confident about the answer] vs. forced generation [i.e., generate whether certain or not] vs. read only). We were able to address the effects of committing errors with this design because in the free generation condition we expected that there would be few errors, whereas in the forced generation condition we expected many errors. We could also address the effect of generation by comparing the generation conditions to the read-only condition. The second factor was feedback. On half of the trials in each generation condition, feedback, which consisted of showing the correct answer for 2.5 sec, was given after each response. On the other half of the trials in these conditions, no feedback was given. This manipulation allowed us not only to investigate the effect of feedback per se, but also to see whether or not feedback had a particularly important effect when the participants were committing many errors. Perhaps feedback is important when errors need to be corrected, but not otherwise. A comparison of the effects of feedback in the free and forced generation conditions would allow us to find out. In the read-only condition, the items were presented and no feedback was given.

Participants. We conducted the experiment first with 16 Grade 6 children attending an at-risk public middle school in Bronx, New York (using as teaching materials definitions such as *To discuss something in order to come to an agreement:* Negotiate) and then with 52 Columbia University students (using more difficult materials—e.g., *Disdainful; characterized by haughty scorn:* Supercilious). We then replicated the experiment with a different group of 26 Columbia University students, allowing them unlimited time to study the feedback. The effects were qualitatively the same in each case, so we will present them together here.

Procedure. The procedure consisted of three phases: (1) initial study, during which all of the definition–word pairs were presented; (2) the manipulation phase described above; and (3) a cued-recall test, in which the participants tried to produce the word corresponding to each of the definitions. Each session was split into two blocks, with one block run under the free generation condition and the other under the forced generation condition. The three feedback conditions (generate with feedback, generate without feedback, and read only) were mixed within each block. Since the read-only condition did not involve generation, it was the same in both blocks.

Results and Discussion

Generation. To our surprise, we found that the generation conditions did not result in memory consistently superior to that of the read-only condition.

Errors. We found no effect of errors. The results in the forced generation condition were not different from those of the free generation condition in any of the three replications of the experiment.

Feedback. In each replication, we found a large and significant effect of feedback. Performance was better when feedback was given than when it was not, regardless of whether generation was free or forced, in every replication of the experiment. This advantage for items that were corrected when the wrong answer (or no answer) was given persisted even when a second study–test trial was given. The data showing the full design, collapsed over all three replications, are shown in Figure 1. As can be seen, only feedback mattered, but it mattered a great deal.

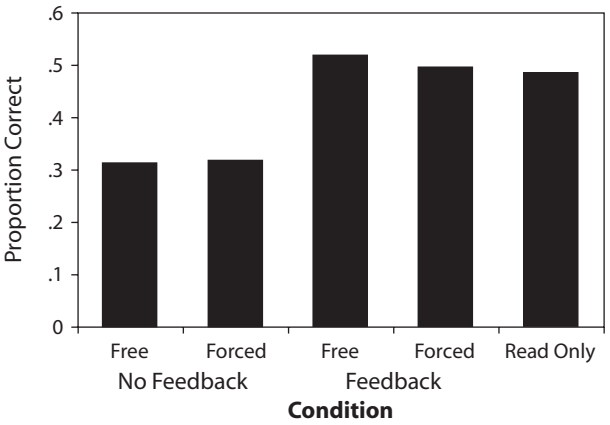

Figure 1. Proportion correct averaged over the three replications of the experiment in generation, errors, and feedback as a function of condition. The results showed no differences among the free condition with feedback, the forced generation condition with feedback, and the read-only control condition. In the absence of feedback, both forced and free generation resulted in significantly poorer memory than in the other three conditions.

FOLLOW-UP STUDY AND IMPLICATIONS

Self-Generation

The failure to find a generation effect came as a surprise to us. Many previous experiments have reported generation effects (e.g., Slamecka & Graf, 1978) as well as boundary conditions (Hirshman & Bjork, 1988; McDaniel, Waddill, & Einstein, 1988), and we were highly confident in the beneficial effects of self-generation. But confidence can be unwarranted. Interestingly, deWinstanley and Bjork (2004) found a generation effect on the first trial but not on the second trial. They interpreted this finding as indicating that by the second trial their participants had learned that it was efficacious to generate the answers and did so even in the read condition. Perhaps, similarly, in our experiment, the participants were in fact generating when they had nominally been assigned to the read-only condition. If so, it would explain the ostensible absence of a generation effect.

We therefore examined our own procedures more closely. Two factors seemed particularly relevant in the experiment reported in the present study. First, in the read-only condition, we had tried to ensure that the participants would read the cue before they read the target by showing the cue alone for 1 sec before presenting the target. In retrospect, this brief pause may have been enough to lead the participants to generate, even though nominally they were in the read-only condition. Second, we had presented the read and generate items intermixed within a single list. Although previous research has shown generation effects in within-list designs (Slamecka & Katsaiti, 1987), we were concerned, especially given the 1-sec pause, that the participants in our experiment might have attempted to generate on every item.

To further test the idea that the brief cue-alone pause was important in allowing people to generate even in the read-only condition, we conducted a follow-up experiment (Metcalfe & Kornell, 2007). In the read-only condition, a cue and a target were presented together for 6 sec simultaneously, whereas in the quasi-generation condition a cue word was presented alone for 3 sec and then remained visible while a target was added for another 3 sec. Recall was

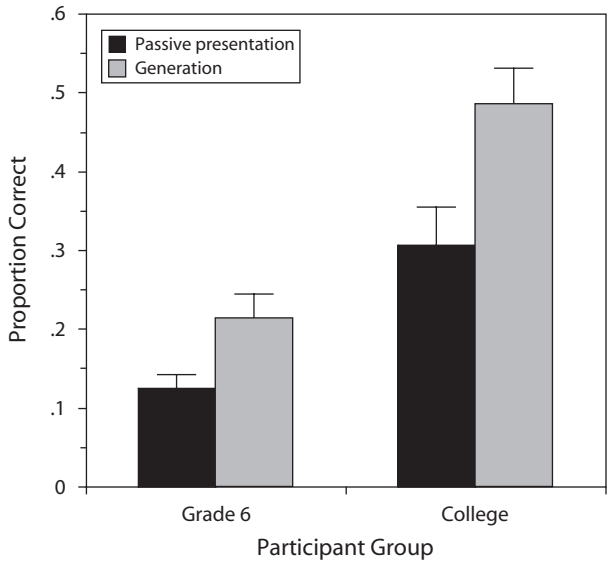

Figure 2. Proportion correct when items were presented in the passive read-only condition versus in the active generation condition for children in Grade 6 (left) and college students (right). Bars represent standard errors of the means.

enhanced by presentation of the cue alone prior to presentation of the target. Carrier and Pashler (1992) reported similar results. Thus, the small and seemingly innocuous pause that we had included only to be sure that our participants paid attention to the cue may have induced them to generate the answer even in the control condition.

But, having failed to replicate the much vaunted generation effect three times, we decided to verify whether or not, when there was no pause and no intermixing of the generate and read items, we would replicate the generation effect. In one such experiment (see Metcalfe & Kornell, 2007), conducted with children in Grade 6 in the Bronx and with students at Columbia University, we found large and significant generation effects, shown in Figure 2.

The important pedagogical implication of the absence of a generation effect in our three replications of the present experiment, as well as in deWinstanley and Bjork's (2004) research, is that it might be easy for a savvy educator to produce a generation effect and thereby greatly enhance his or her students' learning. The read condition in our follow-up experiments, in which a large generation effect was manifested, was such that the learner could passively read (or maybe even not read) the materials. We are inclined to suppose that in a practical situation such a method of presentation may encourage the learner's attention to flag. Rather than spoon-feed students by laying everything out for them, it would behoove the instructor to pause from time to time, asking questions that encourage the students to generate the answers themselves, Socratic style. It would seem important not to immediately jump to the answer. A short pause or, even better, a seemingly interminable pause of several seconds may greatly enhance learning. It may also be beneficial for the whole class, and not just for the students who answer, if the instructor asks individuals to answer aloud, but without saying in advance who the victim will be. Under such conditions—which are rather like our mixed condition, in which the question–pause technique was used—we predict that everyone in the class will start generating all the time, and memory will be greatly improved.

Errors

In the three replications of the first experiment (one conducted with Grade 6 children and two with Columbia students), we found no difference in eventual recall as a function of whether the participants were forced to generate their answers or free to produce only those answers about which they were highly confident. In the first experiment, the participants in the free generation condition produced error rates of 22.1%, 15.4%, and 15.2% in the first, second, and third replications, respectively; in contrast, those in the forced generation condition had error rates of 80.1%, 69.0%, and 70.4%, in the three respective replications. Clearly, there were large differences in how many errors were produced, but equally clearly, when the participants were given corrective feedback on these errors, there was no difference whatsoever in their final correct recall (see Figure 1).

There are reasons to believe that errors *should* have had a negative effect. When a person commits an error, should it not put him or her into an A–B/A–C interference situation, where C is the correct response and B is the error? Should the error not produce the kind of dilemma that people face when given misinformation (Loftus & Hoffman, 1989)? One would expect errors to interfere with memory for the correct items. Indeed, one might expect them to be even more harmful than the responses in the classic interference-theory paradigm or the misleading information paradigm, in which the nontargeted information is merely presented to the participant. When people *generate* their own errors, the errors should be even more memorable—and thus detrimental—than information that is merely presented.

Because this logic seemed compelling, we conducted two additional experiments (Metcalfe & Kornell, 2007), one with middle school children and one with college students, in which participants were instructed either to answer every question (and by necessity make errors) or to avoid errors at all costs. Neither experiment showed an effect of errors. Despite many efforts by now, we have never been able to obtain a detrimental effect of producing an error, as long as corrective feedback is given.

But, one may argue, perhaps one has to *believe* in an error in order for it to cause a detrimental memory effect. If you produce an error that you know is wrong (because the computer program forces you to write *something*), perhaps it should not count as an error at all. Perhaps only errors committed with high levels of confidence have a detrimental effect on memory.

Although this sounds plausible, the data argue against it. Butterfield and Metcalfe (2001; also see Butterfield & Metcalfe, 2006) asked people to give confidence ratings immediately following their answers to general information questions. If only high-confidence errors—that is, errors that people think are correct responses—have

detrimental effects, then these should be the most difficult to correct. In contrast, Butterfield and his colleagues (e.g., Butterfield & Metcalfe, 2001, 2006) have consistently shown that high-confidence errors are the easiest to correct—a finding they refer to as the *hypercorrection effect*. They have investigated the reasons for, and the neural correlates of, the high-confidence error hypercorrection effect, and, irrespective of the reasons for the effect, they have now replicated the same basic finding many times. According to our data, then, errors, when corrected, do not result in interference, and according to the results of Butterfield and his colleagues, errors that people believe in most are corrected even more easily than those about which they are unsure.

Where does this leave us pedagogically? Our results indicate that having students generate is a good idea and that errors do not harm learning as long as they are corrected. The fear that students might generate an error that could be detrimental to their learning the correct answer appears to be unfounded. It should not, according to our data, be used as a reason to keep students from actively generating or from participating fully in their own learning.

Feedback

Our results with respect to feedback are straightforward. Without exception, feedback had a large and important beneficial effect on learning. Most often, when no feedback was given, responses that were initially wrong simply stayed wrong. This seems logical and unsurprising (but worth acting on nevertheless), since in a verbal learning situation errors rarely correct themselves spontaneously. We are currently in the process of investigating when feedback should be given and do not yet know the answer. But we do know that feedback helps: It was the single most influential factor in our results, and we strongly recommend its use. To err may be human, but to give corrective feedback is divine.

AUTHOR NOTE

We thank the Institute of Educational Sciences, Department of Education, for Grant CASL R305H060161, and the James S. McDonnell Foundation for Grant CSEP97-58, which supported the research reported here. We are grateful to Lisa K. Son for her help. Correspondence concerning this article should be addressed to J. Metcalfe, Department of Psychology, Columbia University, New York, NY 10027 (e-mail: jm348@columbia.edu).

REFERENCES

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, **61**, 213-238.

Benjamin, A. S., & Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory & Language*, **55**, 126-137.

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.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, **65**, 245-281.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1491-1494.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition & Learning*, **1**, 69-84.

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.

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

Dunlosky, J., Hertzog, C., Kennedy, M. R. F., & Thiede, K. W. (2005). The self-monitoring approach for effective learning. *International Journal of Cognitive Technology*, **10**, 4-11.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, **81**, 392-399.

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 484-494.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior*, **10**, 562-567.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory & Language*, **35**, 157-175.

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.

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

Loftus, E. F., & Hoffman, H. G. (1989). Misinformation and memory: The creation of new memories. *Journal of Experimental Psychology: General*, **118**, 100-104.

McDaniel, M. A., Waddill, P. J., & Einstein, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory & Language*, **27**, 521-536.

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

Metcalfe, J. (2006). Principles of cognitive science in education. *APS Observer*, **19**, 27-28.

Metcalfe, J., & Finn, B. (in press). Judgments of learning are directly related to study choice. *Psychonomic Bulletin & Review*.

Metcalfe, J., & Kornell, N. (2007). *The effects of generation, errors, and feedback on learning*. Manuscript in preparation.

Metcalfe, J., Kornell, N., & Son, L. K. (in press). A cognitive-science-based program to enhance study efficacy in a high- and low-risk setting. *European Journal of Cognitive Psychology*.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 3-8.

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

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, **3**, 207-217.

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

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory & Language*, **26**, 589-607.